Interim Report

LGP Discrimination and Residual Risk Analysis on Standardized Test Sites—Camp Sibert

ESTCP Project MM-0811

JUNE 2010

Frank D. Francone
RML Technologies, Inc.

Dean A. Keiswetter
Larry M. Deschaine
SAIC

Environmental Security Technology Certification Program

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 2010		**00-00-2010 to 00-00-2010**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
LGP Discrimination and Residual Risk Analysis on Standardized Test Sites - Camp Sibert	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Science Applications International Corp (SAIC),1710 SAIC Drive,McLean,VA,22102**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**131**	

TABLE OF CONTENTS

TABLE OF FIGURES

LIST OF ACRONYMS

-2LL	Minus two times Log Likelihood
11X	Depth corresponding to 11 times an object's diameter
AUC	Area under the Curve of a ROC curve
BRAC	Base Realignment and Closing
CFS	Correlation-Based Feature Selection
CNG	California National Guard
CWS	Chemical Warfare Service
DAQ	Data Acquisition Computer
DGM	Digital Geophysical Mapping
DoD	Department of Defense
EE/CA	Engineering Evaluation/Cost Analysis
EM61MTADS	EM61 MKII MTADS Array
EMI	Electromagnetic Induction
ESTCP	Environmental Security Technology Certification Program
FPU	Floating Point Unit
FUDS	Formerly Used Defense Site
GEMTADS	GEM-3 MTADS Array
GLRT	Generalized Likelihood Ratio Test
GPO	Geophysical Prove-Out
GPS	Global Positioning System
GSA	General Services Administration
HRR	Historical Records Review
IDA	Institute for Defense Analyses
IMU	Inertial Measurement Unit
LGP	Linear Genetic Programming
MAGMTADS	Magnetometer MTADS Array
MEC	Munitions and Explosives of Concern
MSEMS	Man Portable Simultaneous EMI and Magnetometer System
MRMR	Maximum Relevance Minimum Redundancy
MRS	Munitions Response Site

MTADS	Multi-sensor Towed Array Detection System
Nfa	Number of False Alarms
NOSLN	No On Site Learning Necessary
NRL	Naval Research Laboratory
Pclass	Probability of Correct Classification
PDF	Probability Density Function
RTK	Real Time Kinematic
QA	Quality Assurance
QC	Quality Control
RML	RML Technologies, Inc.
ROC	Receiver Operating Characteristic
SAIC	Science Applications International Corporation
SCORR	Slope Corrected
SERDP	Strategic Environmental Research and Development Program
SI	Site Investigation
SLO	San Luis Obispo
SNR	Signal to Noise Ratio
TEMTADS	Time Domain EM Discrimination Array
TOI	Targets of Interest
UTC	Universal Coordinated Time
UXO	Unexploded Ordnance

LIST OF TABLES

EXECUTIVE SUMMARY

The demonstration described in this report was conducted at the Former Camp Sibert, Alabama, under project ESTCP MM-0811 "LGP Discrimination and Residual Risk Analysis at Camp Sibert." It was performed under the umbrella of the ESTCP Discrimination Study Pilot Program. The MM-0811 project demonstrates the application of the LGP Discrimination Process™ to the problem of UXO discrimination.

At the Camp Sibert site the objective was to discriminate potentially hazardous 4.2" mortars from non-hazardous shrapnel, range and cultural debris. Digital Geophysical Mapping ("DGM") was acquired by the ESTCP Program Office from a variety of sensor arrays.

The LGP Discrimination Process™ begins with the DGM from a site suspected of containing UXO. It then extracts attributes from anomalous regions (targets) in the DGM, uses Linear Genetic Programming ("LGP") and the extracted attributes to rank the targets in their order of likelihood of being UXO, and finally, applies statistical residual risk analysis to determine which of the ranked targets may be safely left in the ground as not-UXO.

In this report, we describe the performance of the LGP Discrimination Process in three separate tracks. The tracks were different in the sensor data combinations and the techniques used for attribute extraction. The three tracks may be described as follows:

1. **"EM-only-track."** The sensor set used was the MTADS EM61 Array ("EM61MTADS"). The attributes extracted were statistical attributes drawn from the DGM of that sensor;
2. **"Combined-track."** The sensor sets used were the EM61MTADS and the MTADS Magnetometer Array ("MAGMTADS"). The attributes extracted were statistical attributes drawn from the DGM of both of these sensors;
3. **"Inversion-track."** Under MM-1505, SAIC had previously extracted phenomenological features from the EM61MTADS and MAGMTADS sensors in the Camp Sibert DGM. The attributes used were those phenomenological features.

On all three tracks, the attributes extracted were analyzed by information-theoretic and statistical methods to reduce the attribute set to a small number of highly-predictive attributes. Then, Linear Genetic Programming was used to rank the targets as either UXO or Not-UXO using a small "training" set of targets for which groundtruth was provided. Finally, statistical residual risk analysis was applied to the rankings and to the training groundtruth to determine the stop-digging cut-off.

Predictions on a much larger "blind" data set containing one-hundred and nineteen seeded 4.2" mortars provided the metric for success. On all three tracks, 100% of the UXO were located with only a small number of false-positives and a near-perfect ROC curve was generated by the LGP-generated rankings. On all three tracks, a high percentage of non-UXO were safely left in the ground as high-probability Not-UXO.

The main difference between the performance on the three tracks was in the cannot-analyze targets. For various reasons, a portion of targets on each track had to be classified

as not containing proper data to be safely left in the ground as not-MEC. By that metric, the EM-only-track showed by far the best performance and the Inversion-track the worst.

There appeared to be no advantage to adding Magnetometer data to EM61 data in enhancing discrimination performance for these data. The ROC curves generated by the EM-only-track and the Combined-track were statistically indistinguishable and the addition of the second sensor set required more targets to be classified as cannot-analyze. Thus, overall the discrimination on the EM-only-track was better, measured by percentage on non-UXO left in the ground.

Finally, the intention in this project was to test an iterative process in which the first LGP rankings and risk analysis would be used to select further groundtruth. That further groundtruth would be used as the basis for additional LGP ranking and risk analysis. That process would have iterated until a stop-digging decision was reached. The goal of iteration was to improve the ROC charts and to improve the accuracy of the stop-digging cutoff with additional groundtruth.

We were unable to demonstrate this process because the ROC charts on our first iteration for all three tracks were near-perfect and the stop-digging thresholds accurately identified all UXO. Thus, there was little or no room for improvement with subsequent iterations.

1 INTRODUCTION

1.1 BACKGROUND

The FY06 Defense Appropriation contains funding for the "Development of Advanced, Sophisticated Discrimination Technologies for UXO Cleanup" in the Environmental Security Technology Certification Program. In 2003, the Defense Science Board observed: "The … problem is that instruments that can detect the buried UXOs also detect numerous scrap metal objects and other artifacts, which leads to an enormous amount of expensive digging. Typically 100 holes may be dug before a real UXO is unearthed! The Task Force assessment is that much of this wasteful digging can be eliminated by the use of more advanced technology instruments that exploit modern digital processing and advanced multi-mode sensors to achieve an improved level of discrimination of scrap from UXO."[1]

Significant progress has been made in discrimination technology. To date, these technologies have primarily been tested at constructed test sites, with only limited application at live sites. The routine implementation of discrimination technologies will require demonstrations at real UXO sites under real world conditions.

1.2 OBJECTIVE OF THE DEMONSTRATION

Our objective is to advance and improve MEC discrimination performance by validating a decision process that (i) combines statistical analyses of digital geophysical mapping products and Linear Genetic Programming (LGP) methods to enable classification, and

[1] *Report of the Defense Science Board Task Force on Unexploded Ordnance.* Department of Defense. December (2003).

(ii) provides iterative quantitative residual risk assessments that may be used during the excavation phase to determine a stop-digging cutoff.

1.3 REGULATORY DRIVERS

Senate Report 106-50, pages 291–293, accompanying the *National Defense Authorization Act for Fiscal Year 2000* (Public Law 106-65),[2] included a provision entitled "Research and development to support unexploded ordnance clearance, active range unexploded ordnance clearance, and explosive ordnance disposal." This provision requires the Secretary of Defense to submit to the Congressional defense committees a report that gives a complete estimate of the current and projected costs, to include funding shortfalls, for UXO response at active facilities, installations subject to base realignment and closure (BRAC), and formerly used defense sites (FUDS).

In 2001, the Department of Defense ("DoD") reported to Congress:

> "Decades of military training, exercises, and testing of weapons systems has required that we begin to focus our response on the challenges of UXO. Land acreage potentially containing UXO has grown to include active military sites and land transferring or transferred for private use, such as BRAC sites and FUDS. DoD responsibilities include protecting personnel and the public from explosive safety hazards; UXO site cleanup project management; ensuring compliance with federal, state, and local laws and environmental regulations; assumption of liability; and appropriate interactions with the public.
>
> "…Through limited experience gained in executing these activities, it has become increasingly clear that the full size and extent of the impact of sites containing UXO is yet to be realized. … DoD has completed an initial baseline estimate for UXO remediation cost. This report provides a UXO response estimate in a range between $106.9 billion and $391 billion in current year [2001] dollars. …Technology discovery, development, and commercialization offers some hope that the cost range can be decreased. …
>
> "… Objective: Develop standards and protocols for navigation, geo-location, data acquisition and **processing**, and performance of UXO technologies.
>
> "Standard, high quality archived data are needed for optimal data processing of geophysical data, re-acquisition for response activities, quality assurance, quality control, and review by all stakeholders. In addition standards and protocols are required for evaluating UXO technology performance to aid in selecting the most effective technologies for individual sites.

[2] Senate Report 106-50, National Defense Authorization Act for Fiscal Year 2000, May 17, 1999. *Research and Development to Support UXO Clearance, Active Range UXO Clearance, and Explosive Ordnance Disposal*, pp. 291–293.

"Standard software and visualization tools are needed to provide regulatory and public visibility to and understanding of the analysis and decision process made in response activities."[3]

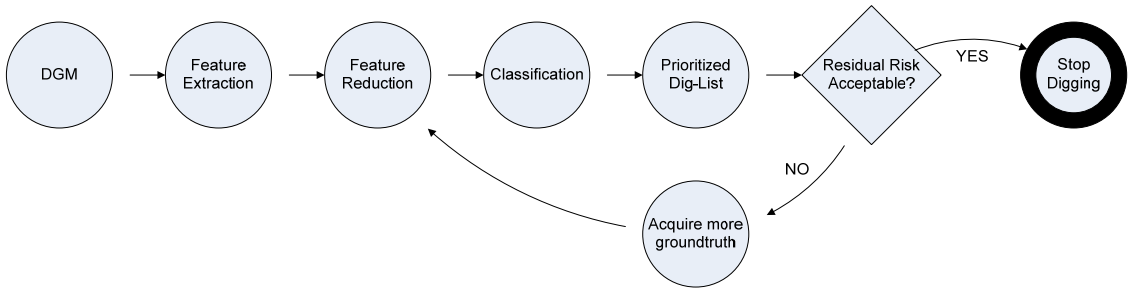
2 TECHNOLOGY

The LGP Discrimination Process is a multi-step, iterative process that uses Linear-Genetic-Programming to perform the most difficult classification tasks and RML's Residual Risk Analysis to recommend a stop-digging decision for a customer-defined threshold. It is typically an iterative process. That is, early classification models are used to select groundtruth on which later models are trained.

When the iterative Residual Risk Analysis is added, the LGP Discrimination Process starts with a small training set for initial prioritization of Targets. If the residual risk is too high to recommend a stop-digging decision, additional ground truth is acquired and that ground truth is added to the training set. From that larger training set a better prioritized dig-list is built. This process continues until reaching a customer designated risk level for the probability that no intact MEC remain on the site.

Figure 1 shows the complete iterative process by which improved classification models are built as the site is excavated. The goal in the iteration is to characterize the tail of the probability density function that a given target is MEC as a function of dig-list ranking with the fewest possible number of excavations. From that tail, the residual risk of MEC remaining on site may be computed to customer specified confidence levels.

Figure 1. The LGP Discrimination Process including iterative residual risk analysis[4]



2.1 TECHNOLOGY DESCRIPTION

The steps in the LGP Discrimination Process are:

1. Data Acquisition
2. Data QAQC
3. Attribute Extraction
4. Attribute Reduction

[3] Department of Defense, *Unexploded Ordnance Response: Technology and Cost*, Report to Congress, March 2001.

[4] We use the term "feature" in this figure to describe what is elsewhere in this report called an "attribute."

5. Modeling
6. Residual Risk Analysis
7. Iterate. Request further Groundtruth and Iterate thru steps 4-6 until stop-digging decision is reached.

We will address each of these steps in this section.

2.1.1 Data Acquisition

The sensors used to collect data for this project were the MTADS magnetometer array ("MAGMTADS") and the MTADS EM61 MKII Array ("EM61MTADS"). The EM61MTADS was configured with an upper and lower coil.

The Digital Geophysical Mapping ("DGM") data from these sensors were provided to us by the ESTCP program office, leveled and lag corrected. So while this is formally a step in our process, we did not perform data acquisition as part of this project.

The DGM generated by the EM61MTADS was used to perform discrimination on the EM-only-track. The DGM generated by the EM61MTADS and the MAGMTADS was used to perform discrimination on the Combined-track.

2.1.2 Data QAQC

The purpose of this step is to assure that the data on which we are performing modeling is good enough to support a no-dig decision for each target. Data QAQC is not a singular step that ends early in the process. It is an ongoing procedure that occurs throughout the LGP Discrimination Process.

Thus, we may determine that the DGM in the region of a target is sufficiently ambiguous or overlapping with an adjacent target that it may not be properly modeled. This would occur toward the beginning of our process. On the other hand, later, and after we have completed the Attribute Reduction step (see below), we observe the resulting distribution of attributes that have been identified as potentially important attributes. Statistical outliers on these attributes would be excluded from further analysis. Finally, after we perform residual risk analysis, we examine attribute space and may determine that the data density in attribute space is not sufficient to support a no-dig decision for a particular target.

The result of this process is that a certain portion of targets are assigned to a "cannot-analyze" category, which means that these targets must be dug regardless as they cannot be confidently designated as high-confidence Not-UXO.

Because of the differing nature of our input data on the three tracks, the procedures for QAQC varied from track to track substantially. Thus, we will address the specifics of QAQC in the portion of this report that addresses each step.

2.1.3 Attribute Extraction

The purpose of attribute extraction is to measure aspects of each target in a way that is meaningful to the ranking of UXO vs. Not-UXO. We use numeric attributes in that regard. Some of those attributes become inputs to the LGP modeling algorithm.

The attribute extraction process was very different as between the three tracks: in the EM-only-track, we worked off of EM61 DGM only. In the Combined-track, we worked off of both EM61 and Mag DGM. In the Inversion-track, we worked off of phenomenological attributes previously extracted by SAIC in MM-0210. Accordingly, we address here, each of those tracks separately.

2.1.3.1 Inversion-Track.

Phenomenological attributes for the Inversion-track were provided to us by SAIC, having been previously extracted by SAIC in MM-0210. In that project, model-based estimation was used to determine parameters of an unknown target, assuming the flux originates from an induced dipole model at the target location. Fitted model parameters include anomaly size (based on the moment for magnetic data and the trace of the polarizability tensor for EMI), shape (EMI only), XY position, depth, orientation, and fit error statistics.[5] These model parameters were used as the attributes for the Inversion-track.

2.1.3.2 EM-only-Track and Combined-Track

Attribute extraction was similar for the EM-only and Combined-tracks. An ellipse was defined for each target. The ellipse separates the region that comprises signal from the region that comprises background noise. Figure 13 shows EMMTADS target 4 with such an ellipse drawn around it.

In addition to the ellipse, circular rings formed by circles centered at the target pick, each circle being 0.75 meters larger than the next smaller one were defined. Each of these rings is a region around a particular target.

From the EM61MTADS DGM, we extracted:

- The first and second moments were measured in each region for channels 1-3, the sum channel and the top-coil channel. The "sum channel" is the sum of the values in the three lower-coil channels.
- The first and second moments were measured of the ratios of adjacent bottom-coil decay channels
- The first and second moments were measured of the ratios of the top-coil channel to channel 3
- The first and second moments were measured of the ratios of the top-coil channel to the sum channel.

For the Combined-track, we extracted all of the above attributes. In addition, we used the analytic signal generated by Oasis Montaj for the MAGMTADS DGM. That comprises a single channel and we took the first and second moments of the above defined regions around each Mag target.

---

[5] The Interim Report for this aspect of feature extraction provides extensive detail on methodologies used by SAIC for this aspect of our feature extraction on the current project. SAIC, *SAIC Analysis of Survey Data Acquired at Camp Sibert*, ESTCP Project MM-0210. Available at www.ESTCP.org.

2.1.4 Attribute Reduction

The Attribute Extraction process described above produces hundreds of statistics for every target. The goal in attribute reduction is to reduce the number of attributes used in modeling to just a handful of highly relevant attributes that contain complementary information content about the modeling problem.

We used a collection of tools at different points in the modeling process to reduce attributes. The tools include: (1) Numeric Input Binning; (2) Maximum Relevance Minimum Redundancy ("MRMR"); (3) Correlation-Based Feature Selection; (4) Decision Trees; and (5) Discipulus™ Input Impacts analysis.

A more detailed description of these techniques may be found in Section 6.6.

2.1.5 Modeling

Modeling is the process of mapping the subset of attributes created in the attribute selection process to the groundtruth of UXO vs. Not-UXO. Our principal modeling tool is RML's Linear Genetic Programming ("LGP") software, Discipulus™ modified to use Area under the curve as a fitness function.

RML's LGP is an inductive-learning technology that is a variant of canonical Genetic Programming. Learning is conducted on a training dataset, consisting of an n-tuple for each Target, comprised of $n-1$ features that describe the Target and a class-label for the Target. The class-label, for MEC discrimination is, of course, whether the target is or is not MEC.

$r[1] = r[1] + x$
$r[1] = r[1] - 1$
$r[0] = r[1] * r[1]$
$r[1] = r[0] * r[1]$
$Output = r[0] + r[1]$

During training, LGP creates computer functions comprised of very simple Intel Floating Point Unit ("FPU"), machine-code instructions such as +,*,-,/, sqrt, power. Internal computations in the function operate directly on the FPU registers and the $n-1$ input features stored in memory. The LGP-created functions map the $n-1$ features to an output that orders the targets in terms of the likelihood they are MEC.

That ordering results in a prioritized dig-list. A simple five-line LGP function might look like the pseudo-code in the text box. (All registers are represented by $r[n]$ and are initialized to zero. The one input feature in this example is represented by x). This program uses two registers to represent a functional mapping of x to an output, $f(x)$. The function, in this case, is the polynomial, $f(x) = (x-1)^2 + (x-1)^3$.

LGP's learning algorithm has been described in detail in the literature.[6] In brief, LGP is a steady-state, evolutionary algorithm using tournament-selection to continuously improve a population of Intel machine-code functions. A single run is comprised of tournaments that compare the "fitness" of two randomly-selected programs that are repeated until a

---

[6] Banzhaf, W., Nordin, P. Keller, R. Francone, F. (1998) *Genetic Programming, an Introduction*, Morgan Kaufman Publishers, Inc., San Francisco, CA at pp 257-264; and Nordin, J.P., Francone, F., and Banzhaf, W. (1999) "Efficient Evolution of Machine Code for CISC Architectures Using Blocks and Homologous Crossover," in *Advances in Genetic Programming 3*. Chapter 12 (MIT Press, Cambridge MA).

termination criterion is reached. At that point, the Intel machine-code of the selected best functions is decompiled into a readable and understandable C-code function. In practice, LGP is configured to perform many runs sequentially and to optimize its own parameters as those runs proceed.

For smaller training sets, we add noise to the inputs. The amount of noise is defined by a percentage parameter *noise_%*. The larger the *noise_%* parameter, the wider the standard deviation of the added noise. The number of training instances is multiplied by another parameter, each instance having noise added to the inputs.

The *noise_%* parameter is set using k-fold cross-validation.[7]

The LGP models are then trained on data prepared using a technique called bagging, with the *noise_%* set to the previously selected value. Assume a training data set of size *n*. Bagging creates *j* separate training sets. Each training set is prepared by sampling rows from the training set *n* times with resampling.[8] An LGP model is trained from each bootstrap sample and the models are then applied to the blind data. The prediction on for a blind target is the average ranking of each blind target by the multiple LGP models.

2.1.6 Residual Risk Analysis

The final step in each iteration of LGP Discrimination process is RML's Residual Risk Analysis. The goal of Residual Risk Analysis is to recommend a stop-digging decision based on the actual empirical results of applying the LGP Discrimination Process to a particular site, given a customer specified confidence level. The iterative process comprising the Residual Risk Analysis process is shown in Figure 1 and described generally in the text accompanying that figure.

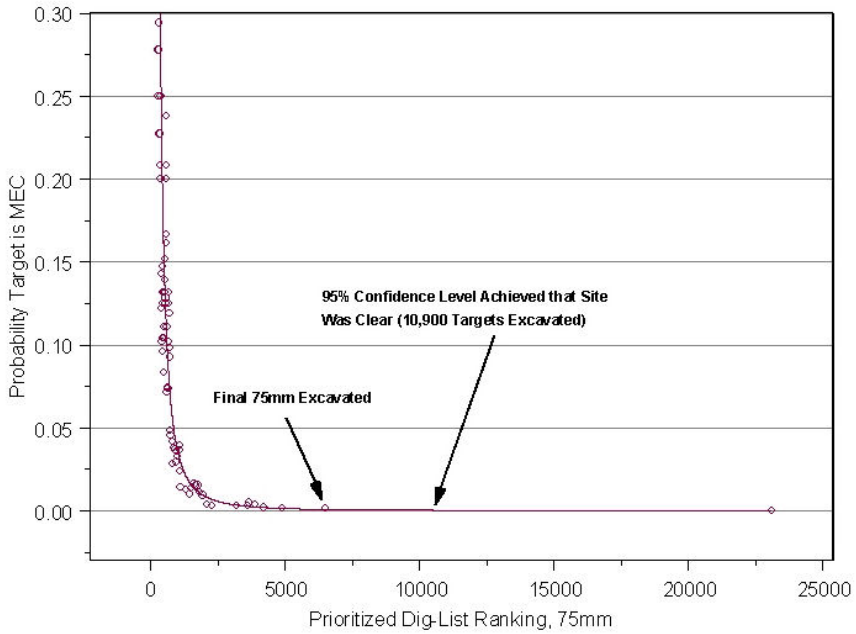
A key property of a prioritized dig-list that accurately discriminates UXO from other items is that the MEC are ranked nearer the start of the dig-list than clutter, hot-rocks, etc. As a result, *as excavation proceeds, the probability that the next item is MEC falls, not always continuously, but it falls.* Figure 2 shows that relationship in our work at F.E.Warren AFB. This is an example of a probability that falls relatively continuously as the dig-list ranking increases.

---

[7] Kohavi, Ron (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2 (12): 1137–1143.
http://citeseer.ist.psu.edu/kohavi95study.html. (Morgan Kaufmann, San Mateo)

8 Breiman, L. (1996). "Bagging predictors". Machine Learning 24 (2): 123–140.

Figure 2. Relationship between prioritized dig-list ranking and probability that a target was 75mm UXO at F.E.Warren AFB.



The circles represent a measured probability that Targets were MEC in the vicinity of each MEC item found. The falling probability of UXO as Rank increases is clearly shown. The red line in Figure 2 is the maximum likelihood power-law relationship fit to these data after ranking 278. (The fit started there because the linear portion of the log-log transformed data in these data started at ranking 278.)

A classifier that produces a high-quality ROC chart will always have the property of falling empirical probability as rank increases. In this step, we fit an appropriate, simple model to the declining probability of UXO as a function of rank. Rank is calculated using the predictive scores output by LGP and the scores are combined across training and blind data to create a common ranking metric for the two data sets. Candidates for the most appropriate model that we considered in this project are Power Law fit, Exponential fit, Kernel Regression fit or Logistic Regression fit.

Once the model is fit on labeled, training data, we predict the probability of UXO as a function of rank for unlabeled, blind data. From those probabilities, we also predict the probability that any sequence of targets from the *nth* ranked target to the maximum ranked target contain one or more UXO. That probability is computed as the OR of the probabilities of UXO for all targets from the nth ranked target to the maximum ranked target. Thus, at any given target ranking, the risk remaining (probability) that the targets with a *higher* ranking (less likely to be UXO) contain one or more UXO items is the or'ed probabilities of all higher ranking targets.

The OR operator when applied to the probabilities of two events labeled A and B (for example, target A or target B being UXO), is computed as follows:

Equation 1:

$$P(A\_OR\_B) = P(A) + P(B) - P(A\_AND\_B)$$ [9]

In the present study, the above formula is applied to all targets ranked to the right of the plotted rank (that is, ranked less likely to be UXO) by chaining the computation. This is applied as follows: Assume that three targets have a higher ranking than a given rank and that the targets are labeled A, B, and C. Given the definition of *P(A_OR_B)* in Equation 1 above, we can now compute the probability of A OR B OR C as follows:

Equation 2:

$$P(A\_OR\_B\_OR\_C) = P(A\_OR\_B) + P(C) - P((A\_OR\_B)\_AND\_C)$$

Equation 2 may be expanded to compute the OR value for the probability that at least one of any number of targets is UXO.

Thus, at any one step, we measure the residual risk using this OR of probabilities computation. The key point here is that the probabilities used in our Residual Risk Analysis are based on the actual, site-specific empirical results of applying the LGP-based dig-list to the site.

2.1.7 Iteration

At each risk analysis step, and based on the ground truth at that time, we estimates the Target parameters described above using the LGP discrimination process (resulting in a prioritized dig-list) and determine if a stop-digging decision is warranted at the specified confidence level. If not, we request more ground truth, re-estimate the parameters using all ground truth then available, and determine (based on the new estimates) if a stop-digging decision is warranted. That process continues until a stop-digging decision is warranted at the specified confidence level.

2.2 TECHNOLOGY DEVELOPMENT

This technology has not been previously developed under grant from ESTCP.

2.3 ADVANTAGES AND LIMITATIONS OF THE TECHNOLOGY

Key differences between LGP and other learning algorithms are:

1. LGP does not just derive parameters for a specified functional form—it derives the functional form itself and optimizes the parameters of the derived functional form, in one pass;
2. Because LGP software operates directly on populations comprised of Intel machine code functions, it is approximately two orders of magnitude faster than comparable inductive-learning technologies.[10] Coupled with the fact that this

[9] Kachigan, S. (1986) *Statistical Analysis*, Radius Press, NY, NY.

[10] Banzhaf, W., Nordin, P. Keller, R. Francone, F. (1998) *Genetic Programming, an Introduction*, Morgan Kaufman Publishers, Inc., San Francisco, CA at pp 257-264; and Nordin, J.P., Francone, F., and Banzhaf, W. (1999) "Efficient Evolution of Machine Code for CISC Architectures Using Blocks and Homologous Crossover," in *Advances in Genetic Programming 3*. Chapter 12 (MIT Press, Cambridge MA); and

software can run on multiple CPU's over a network in parallel, LGP is capable evaluating millions of functions on large data sets in commercially reasonable time-frames;

3. LGP software has been subjected to extensive in-house and third-party testing on a wide variety of data sets over a nine-year period. Results have been published by RML and SAIC[11] and by third-parties[12];
4. LGP was designed to prevent, insofar as possible, building models of the training-set noise rather than the signal sought to be modeled. (LGP's resistance to fitting noise has been noted in the literature; and
5. The version of Discipulus used in this project uses as its fitness function, the area under the curve ("AUC") of the ROC curve defined by the evolved program ranking. In other words, the evolution process is geared toward creating a good ranking. Most other inductive learning algorithms perform some kind of classification and then convert that into a ranking. This is a subtle but important difference because classifying items as, say, UXO vs. not-UXO is a different goal than ranking them well. Discipulus produces much better rankings when it uses an AUC fitness function than it does when using a classification fitness function.

A disadvantage of LGP is that it requires experienced data modelers for its operation. It is a very powerful modeling tool because of the breadth of the search it can conduct over a very large solution space—both because of its speed and because it evolves functional form, not just parameterization of a preexisting functional form. If used improperly, it can produce wonderful-looking results on known data and very poor results when applied to new data.

Fukunaga, A., Stechert, Mutz, D. (1998) "A Genome Compiler for High Performance Genetic Programming," in: *Proceedings of the Third Annual Genetic Programming Conference*, Jet Propulsion Laboratories, California Institute of Technology Pasadena, CA, Morgan Kaufman Publishers, pp. 86–94.

[11] Several years of comparative studies by RML and SAIC are reported in: Francone, F. D., and Deschaine, L.M., (2004) *Extending the Boundaries of Design Optimization by Integrating Fast Optimization Techniques with Machine-Code-Based Linear Genetic Programming, Information Sciences Journal—Informatics and Computer Science*, Elsevier Press, Vol. 161/3-4 pp 99-120 (see sections 8.3-8.6 for results of the comparative study) Amsterdam, the Netherlands. In brief summary, RML's LGP software consistently performs as well as the best-tested alternative classification algorithms or better, on blind data. Other learning algorithms sometimes perform as well as RML's LGP algorithm but are not nearly as consistent as RML's LGP in producing high-quality results on unseen, testing data.

[12] See: (1) Mukkamala, S., Sung, A., Abraham, A., (2004) "Modeling Intrusion Detection Systems Using Linear Genetic Programming Approach," in *Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*; (2) S. Mukkamala, Q. Liu, R. Veeraghattam, A. H. Sung (2005) "Computational Intelligent Techniques for Tumor Classification (Using Micro array Gene Expression Data)." *International Journal of Lateral Computing*, Vol.2, No. 1, ISSN 0973-208X, pp. 38-45; and (3) Mukkamala, G. D. Tilve, A. H. Sung, B. Ribeiro, A . S. Vieira (2006) Computational Intelligent Techniques for Financial Distress Detection. *International Journal of Computational Intelligence Research.*

3 PERFORMANCE OBJECTIVES

The relevant objectives include (i) Target-of-Interest retention rate, (ii) non-Target-of-Interest reduction rate, and (iii) analysis time. The focus will be on identifying items that may be safely left in the ground. The main failure is misclassifying a target of interest as an item that can be left in the ground.

Items that may be safely left in the ground shall include HE fragments, single fins, cultural debris and geology.

Table 1. Performance objectives summary

Performance Objective	Metric	Data Required	Success Criteria	Result
Target-of-Interest retention rate	Percent Target-of-Interest correctly classified as Target-of-Interest at demonstrator stop-digging recommendation	1. Prioritized dig-list 2. Excavation results or scoring report	>0.95	Success
Non-Target-of-Interest reduction rate	Number of false targets eliminated at demonstrator stop-digging recommendation	3. Prioritized dig-list 4. Excavation results or scoring report	>40%	Success
Analysis time	Person-days in production until stop-digging recommendation	5. Log of data analysis time	< 60 person-days	Success on two tracks. Failure on one track.

The following sections provide a more detailed description of these objectives.

3.1 OBJECTIVE: TARGET-OF-INTEREST RETENTION RATE

The effectiveness of the technology for discrimination of munitions is a function of the degree to which responses that do not correspond to targets of interest can be eliminated with high confidence while retaining Targets-of-Interest. This objective measures the retention rate of Targets-of-Interest.

3.1.1 Metric

Compare the number of 4.2”mortars that were correctly classified as of the stop-digging recommendation to the total number of 4.2” mortars detected.

3.1.2 Data Requirements

The data requirements are straightforward; namely, our (i) prioritized dig-list; (ii) our stop-digging recommendation; and (iii) ground truth information.

3.1.3 Success Criteria

The objective will be considered to be met if more than 95% of the Targets-of-Interest are retained after classification.

3.1.4 Result

100% of Targets-of-Interest were retained after classification on all three tracks.

3.2 OBJECTIVE: NON-TARGET-OF-INTEREST REDUCTION RATE

The effectiveness of the technology for discrimination of munitions is a function of the degree to which responses that do not correspond to targets of interest can be eliminated with high confidence while retaining Targets-of-Interest. This objective measures our ability to reduce false alarms.

3.2.1 Metric

Compare the number of non-Targets-of-Interest that were correctly classified as non-Targets-of-Interest to the total number of non-Targets-of-Interest originally detected.

3.2.2 Data Requirements

The data requirements are straightforward; namely, our (i) prioritized dig-list; (ii) our stop-digging recommendation; and (iii) ground truth information.

3.2.3 Success Criteria

The objective will be considered to be met if more than 40% of the non-Targets-of-Interest would remain unexcavated, given our stop-digging decision.

3.2.4 Result

This objective was met on all three tracks. The percent of non-Target-of-Interest remaining in the ground at the completion of the project on the three tracks is shown in

Table 2. Percent of non-Targets-of-Interest remaining in ground

Track	Percent non-Target-of-Interest Left in Ground
EM	89.6%
EM MAG Combined	86.8%
Inversion	67.1%

3.3 OBJECTIVE—ANALYSE TIME AND COST

3.3.1 Metric

Person-days-in-production until stop-digging recommendation. Combined with the daily analysis costs of the production costs of this technology, this gives the per-anomaly cost.

3.3.2 Data Requirements

Days-in-production will be determined from a review of the analyst's time logs and computer run times.

3.3.3 Success Criteria

For this initial demonstration, and given the constraints of data set size and the new data formats involved, the objective will be considered to be met if the stopping criterion is reached in no more than 60 person-days of production time.

3.3.4 Result

This project succeeded on the Combined and Inversion-tracks. It failed on the EM-only-track.

The approximate man-days spent per track in production of the reported results are:

1. EM-only-track: 74.5 man-days;
2. Combined-track: 52 man-days;
3. Inversion-track: 23 man-days.

4 SITE DESCRIPTION

4.1 SITE SELECTION

ESTCP selected Camp Sibert as the demonstration site. Camp Sibert is located within the boundaries of Site 18 of the former Camp Sibert FUDS. The land is under private ownership and is used as a hunting camp.

The criterion that drove the site selection process were (i) a single use artillery or mortar range, (ii) simple clutter environment, (iii) benign geology, (iv) live ordnance used, and (v) benign topography and vegetation. Additional considerations were size (20-25 acres was desired), anomaly density (mostly isolated anomalies; 100-200 per acre), total anomaly count (2,500 to 5,000 anomalies were desired), and access/authorization to seed site with inert targets.

4.2 SITE HISTORY

The former Camp Sibert is located in the Canoe Creek Valley between Chandler Mountain and Red Mountain to the northwest, and Dunaway Mountain and Canoe Creek Mountain to the southeast. Camp Sibert is comprised of mainly sparsely inhabited farmland and woodland and encompasses approximately 37,035 acres. The City of Gadsden is growing towards the former camp boundaries from the north. The Gadsden Municipal Airport occupies the former Army airfield in the northern portion of the site. The site is located approximately 50 miles northwest of the Birmingham Regional Airport or 86 miles southeast of the Huntsville International Airport. The site is near exit 181 off of Interstate 59 in Gadsden and located approximately 8 miles southwest of the City of Gadsden, near the Gadsden Municipal Airport.

Camp Sibert was acquired in July 1942 by the U.S. Army as a replacement training center for the Chemical Warfare Service (CWS). The second Chemical Warfare School was also established there during World War II. At Camp Sibert the CWS conducted various training exercises such as smoke screen defense, chemical decontamination, chemical depot maintenance, and chemical impregnation of clothing. Chemical troops equipped the camp with chemical field filling stations, a toxic gas yard, and decontamination areas. The Army also constructed an airfield for simulation of chemical air attacks against the troops. The camp was closed at the end of the war in 1945, and the chemical school transferred to Ft McClellan, Alabama. The Army declared the property excess and transferred it to the War Assets Administration on 18 November 1946, and then to the Farm Mortgage Corporation. The government terminated the leases on the area on 13 December 1946. After decontamination of the various ranges and toxic areas in 1948, the land was transferred back to private ownership. The airfield, however, was transferred to the City of Gadsden.

4.3 MUNITIONS CONTAMINATION

The munitions-of-concern at Camp Sibert is a 4.2" mortar.

5 TEST DESIGN

The demonstration used MTADS Magnetometry and MTADS EM61 MkII array data acquired at Camp Sibert as part of the ESTCP UXO Discrimination Pilot Program. Details of the MTADS acquisition systems and plans are presented in Technology Demonstration Plan entitled *MTADS Demonstration at Camp Sibert, Magnetometer / EM61 MkII / GEM-3 Arrays.*[13] A summary of the data collection activities, taken from their report, follows.

5.1 CONCEPTUAL EXPERIMENTAL DESIGN

The magnetometry and EMI data were acquired using standard MTADS data collection procedures. For the EMI array, this included surveying the field twice along transects with perpendicular headings.

5.2 SITE PREPARATION

A Geophysical Prove Out area (GPO) was established near the main demonstration area prior to the main demonstration data collection. The GPO was used to verify the anomaly detection thresholds for the three MTADS sensor systems to be demonstrated in the Study. The other data collection demonstrators also validated their systems and methods using the GPO. The GPO was surveyed with each sensor platform prior to data collection in the main demonstration area with that sensor array. The intent of data collection in the GPO with each system is to verify that the items of interest are detected at the depths of interest under site-specific conditions and to validate the selected detection threshold for each sensor array.

---

[13] ESTCP MM-0533. MTADS Demonstration at Camp Sibert, Magnetometer / EM61 MkII / GEM-3 Arrays, Technology Demonstration Data Report, G.R. Harbaugh, D.A. Steinhurst, N. Khadr, September 26, 2007.

Inert 4.2 inch mortars were emplaced within the survey area.

5.3 SYSTEM SPECIFICATIONS

The MTADS hardware consists of a low-magnetic-signature vehicle that is used to tow the different sensor arrays over large areas (10 - 25 acres / day) to detect buried UXO. The MTADS tow vehicle and magnetometer array are shown in Figure 3. Positioning is provided using high performance Real Time Kinematic (RTK) Global Positioning System (GPS) receivers with position accuracies of ~5 cm. The positioning technology requires the availability of one or more known first-order survey control points.

The MTADS magnetometer array is a linear array of eight Cs-vapor magnetometer sensors (Geometrics, Inc., G-822ROV/A). The sensors are sampled at 50 Hz and typical surveys are conducted at 6 mph; this results in a sampling density of ~6 cm along track with a horizontal sensor spacing of 25 cm. A single GPS antenna placed directly above the center of the sensor array is used to measure the sensor positions in real-time (5 Hz). All navigation and sensor data are time-stamped with Universal Coordinated Time (UTC) derived from the satellite clocks and recorded by the data acquisition computer (DAQ) in the tow vehicle.

Figure 3. MTADS tow vehicle and magnetometer array

The EM61 MkII MTADS array is an overlapping array of three pulsed-induction sensors specially modified by Geonics, Ltd. based on their EM61 MkII sensor with 1m x 1m sensor coils. The sensors employed by MTADS have been modified to make them more compatible with vehicular speeds and to increase their sensitivity to small objects. The timing of the gates has been altered and the delay times are given in Table 3. Differential mode will be used for this demonstration. Nominal survey speed is 3 mph and the sensor readings are recorded at 10 Hz. This results in a down-track sampling of ~15 cm and a cross-track interval of 50 cm. In order to obtain sufficient “looks” at the anomalies, or to insure illumination of all three principle axes of the anomaly with the primary field, data is collected in two orthogonal surveys. The EM61 array being pulled by the MTADS tow vehicle is shown in Figure 4.

Individual sensors in the EM61 MkII array are located using a three-receiver RTK GPS system. An Inertial Measurement Unit (IMU) is also included on the sensor array to provide complimentary platform orientation information. The IMU is a Crossbow VG300 running at 30 Hz. A close-up view of the sensor platform is shown in Figure 5

which shows the three GPS antennae and the IMU (black box under the aft port GPS antenna).

Table 3. NRL EM61 MkII Gate timing parameters

Channel	4 Gate Mode	Delay (μs)	Differential Mode	Delay (μs)
1	Bottom Coil	307	Bottom Coil	307
2	Bottom Coil	508	Top Coil	307
3	Bottom Coil	738	Bottom Coil	738
4	Bottom Coil	1000	Bottom Coil	1000

Figure 4. MTADS EM61 array pulled by the MTADS tow vehicle

.

Figure 5. Close-up of MTADS EM61 array with GPS and IMU

5.4 CALIBRATION ACTIVITIES

The standard performance checks performed by the MTADS crew included three types of measurements. At the beginning of field work and again each morning quiet, static data are collected for a period (10 - 20 minutes) with all systems powered up and warmed up (typically 20-30 minutes). For the EM61 array, a 4" diameter Aluminum (Al) sphere is placed a standard distance above the center of each sensor coil several times in sequence to verify the response of each sensor to each object. The system is stationary for this data

collection. Finally, a systems timing check using a fixed-position wire or chain placed on the ground is conducted.

5.5 DATA COLLECTION PROCEDURES

Nova Research, Inc. conducted three total coverage surveys of the final demonstration site (15 acres, four areas). These surveys were conducted using the Naval Research Laboratory (NRL) Multi-sensor Towed Array Detection System (MTADS) magnetometer, EM61 MkII, and GEM-3 (GEMTADS) arrays. These data were collected in accordance with the overall study demonstration plan including system performance characterization including the use of emplaced calibration items and the installed geophysical prove-out area (GPO).

The data collection performance metrics and production rates are shown in Table 4 and Table 5, respectively.

Table 4. MTADS Performance Objectives/Metrics[14]

Type of Performance Objective	Performance Criteria	Expected Performance (Metric)	Performance Confirmation Method	Actual Performance Objective Met?
Qualitative	*Reliability and Robustness*	*General Observations*	*Operator feedback and recording of system downtime (length and cause)*	*Yes*
Quantitative	*Survey Rate*	*Varies with sensor array, 5 (EM) – 20 (Mag) acres / day*	*Calculated from survey results*	*Yes*
	Data Density	*> 30 pts / m^2*	*Calculated from survey results*	*Yes*
	Percentage of Assigned Coverage Completed	*100% as allowed by topography / vegetation*	*Calculated from survey results*	*Yes*
	Location of Modeled Anomalies	*Horizontal: < ± 0.15 m Vertical: < 30% for depths ≥ 30 cm, < ± 0.15 m depths < 30 cm*	*Comparison of model results to known data on emplaced items or validation data on remediated items*	*Yes*
	Detection of GPO items of interest to depth of interest using determined thresholds	*100%*	*Comparison of anomaly lists from GPO to GPO ground truth for each sensor array*	*Yes*
	Data throughput	*All data QC'ed in real time and results (data and anomaly analysis) provided as required by Program Office*	*Analysis of records kept / log files generated while in the field and recorded delivery times*	*Yes*

[14] ESTCP MM-0533. MTADS Demonstration at Camp Sibert, Magnetometer / EM61 MkII / GEM-3 Arrays, Technology Demonstration Data Report, G.R. Harbaugh, D.A. Steinhurst, N. Khadr, September 26, 2007.

Table 5. Survey rates[15]

Sensor System	Survey Time (hours)	# of Field Days	# of Std. Survey Days	Survey Rate (acres / std. day)
Magnetometer	16.1	2	2.0	7.8
EM61 MkII	36.0	5	4.5	3.5

5.6 VALIDATION

All of the targets selected by the Program Office for analysis and provided to us a blind data were selected for validation. There was no sub-selection.

6 DATA ANALYSIS AND PRODUCTS FOR EM-ONLY-TRACK

The EM61 MTADS only track ("EM61MTADS" Track) used statistical attributes extracted from Camp Sibert EM61MTADS data. The targets included in this track were all targets selected by the Program Office as an EM61MTADS target ("EM Targets").

The steps in the LGP Discrimination Process for this track were:

1. Data QAQC
2. Ellipse Definition
3. Exclude "Cannot-Analyze" targets
4. Attribute Extraction
5. Attribute Reduction
6. Modeling
7. Risk Analysis
8. Prioritized Dig-List

We note up front that this project took, what to us, was a somewhat surprising direction. To wit, because of rut-noise issues in the southwest section of the site, we were required to perform two modeling steps, rather than the one modeling step we had anticipated. For convenience, we refer to these two steps as: (1) the Amplitude Discriminator step; and (2) the LGP Modeling step.

We will describe how we handled that rut noise and the steps described above in the following sections.

6.1 DESCRIPTION OF DATA

We received features for 908 EM Targets. The 908 targets were comprised of:

- 174 training (or "labeled") targets. These were the EM Targets for which we knew ground truth; and

[15] ESTCP MM-0533. MTADS Demonstration at Camp Sibert, Magnetometer / EM61 MkII / GEM-3 Arrays, Technology Demonstration Data Report, G.R. Harbaugh, D.A. Steinhurst, N. Khadr, September 26, 2007.

- 734 blind-data (or “unlabeled”) targets. These were targets for which we did not know ground truth.

The ground-truth for the training data was described as follows by the program office.

Table 6. Training groundtruth for EM-only-track

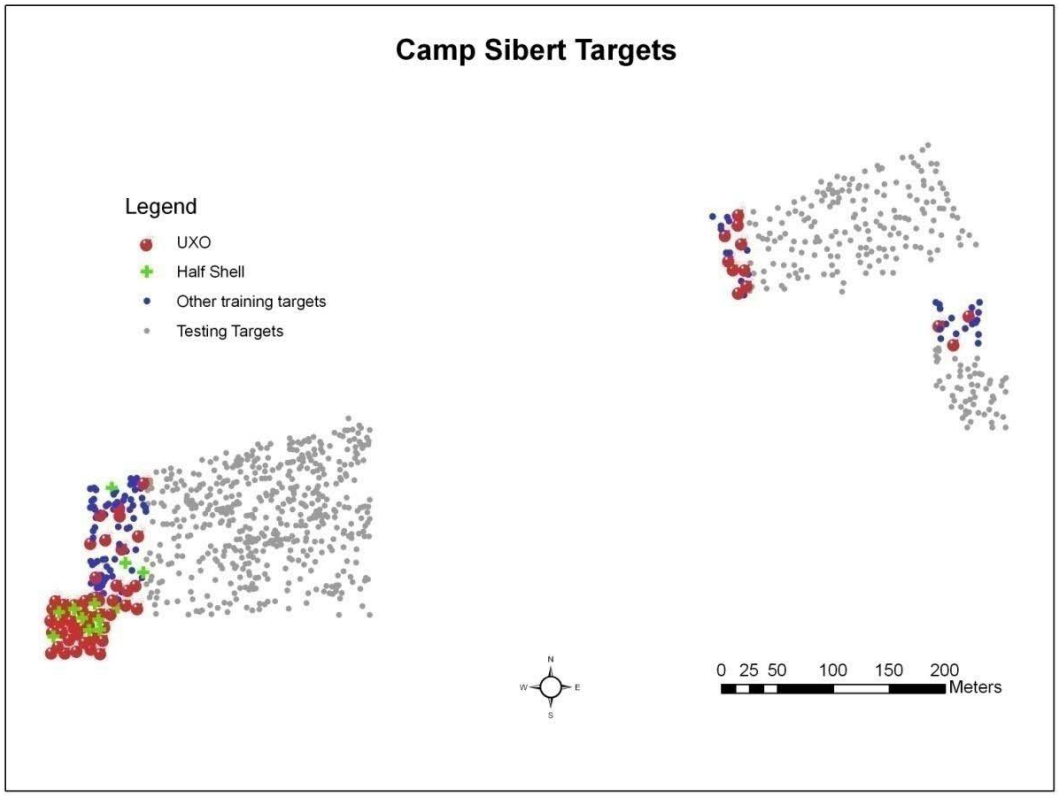
Target Type	*Count*
Baseplate	8
Corner_Stake	3
Frag	14
Halfshell	12
Horseshoe	1
Nose_Frag	9
No_Contact	2
Rock	3
Scrap_Metal	23
Soils	37
Survey_Point	1
UXO	59
Wire	1
Wrench	1
Total	174

The EM digital geophysical mapping (“DGM”) was comprised of 1,724,633 individual data points. Approximately ½ of them were taken on roughly North-South transects and the other half on roughly East-West transects.

The EM data was collected with an EM61MK2 MTADS sensor configured with three decay channels and one lower-coil channel. For convenience, we will refer to the first decay channel as Channel 1, the second as Channel 2 and the third as Channel 3. The top or upper coil reading will be referred to as such. We also summed channels 1-3 into a single channel. We will refer to that as the “sum channel.”

Figure 6 is a spatial map of all targets designated by the program office. Blind data is shown in gray. Training data is colored, depending on what type of munition is reported there in the groundtruth.

Figure 6. Spatial distribution of targets at Camp Sibert testbed.



The site separates into four regions, which we will refer to as follows:

1. **GPO**. The ground prove-out regian is in the lower left hand corner of Figure 6. It appears mostly red and green in color.
2. **Southwest Region**. In the southwest quadrant, all targets outside the GPO are in what we will refer to here as the southwest region.
3. **Northeast Regions**. The two smaller regions in the eastern part of the site will be referred to as the "northeast regions."

6.2 DATA QA/QC AND PREPROCESSING

This section describes the QAQC and Preprocessing used for EM61MTADS data in the EM-only-track and in the EM MAG Combined-track. Although determining which targets cannot be analyzed is fairly part of the QAQC process, we defer discussion of that issue until Section 6.4.

6.2.1 Positional Error

We assessed the quality of the positioning of the data points in two ways: (1) Distance between adjacent points; and (2) Time difference between adjacent points. To do so, we randomly sampled 100,000 pairs of adjacent data points from the DGM. We then measured the difference between their position and between the times at which the

samples were taken. A selection of outliers was selected and a satisfactory explanation was derived for each.

Figure 7 shows the distribution of the point-to-point distances and Figure 8 shows the time differences.

Figure 7. Distance between 100,000 randomly sampled data points. Outliers greater than one meter in distance excluded.

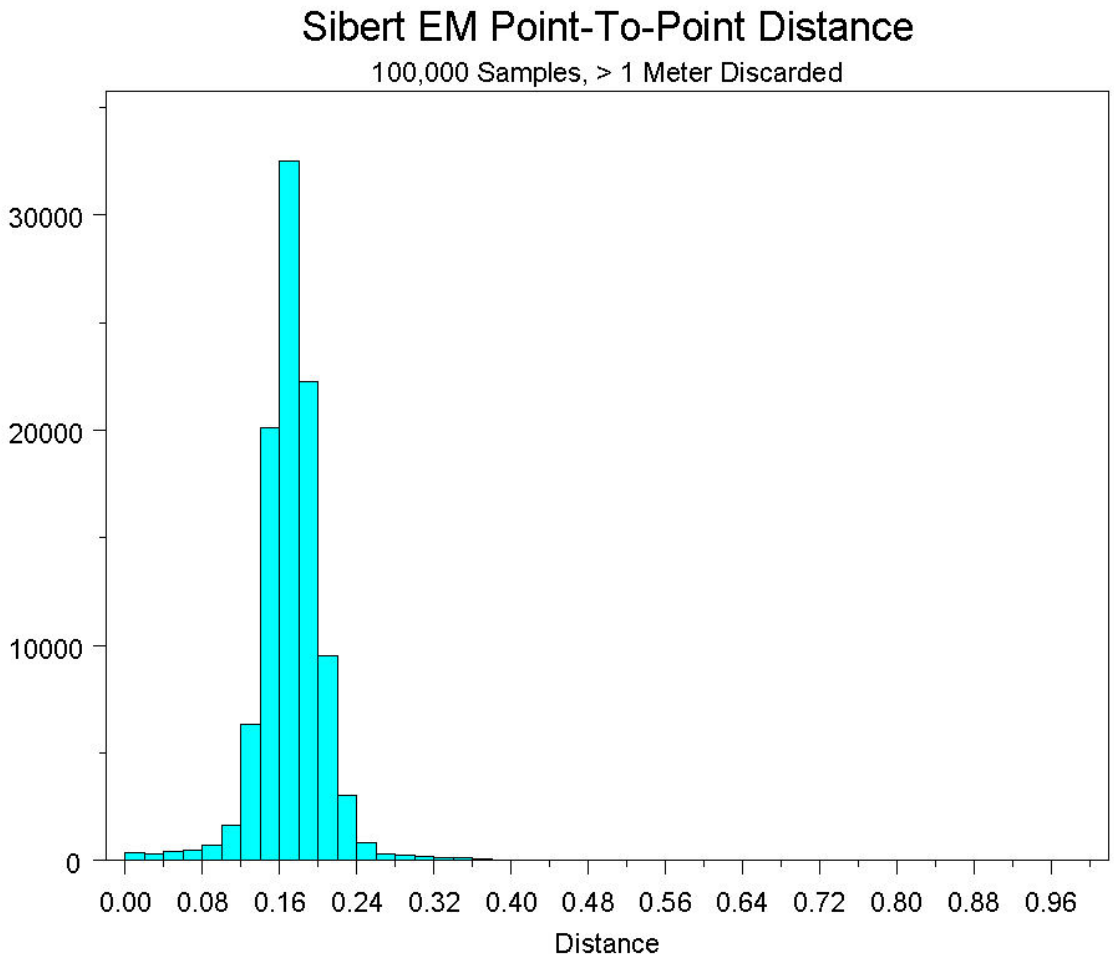
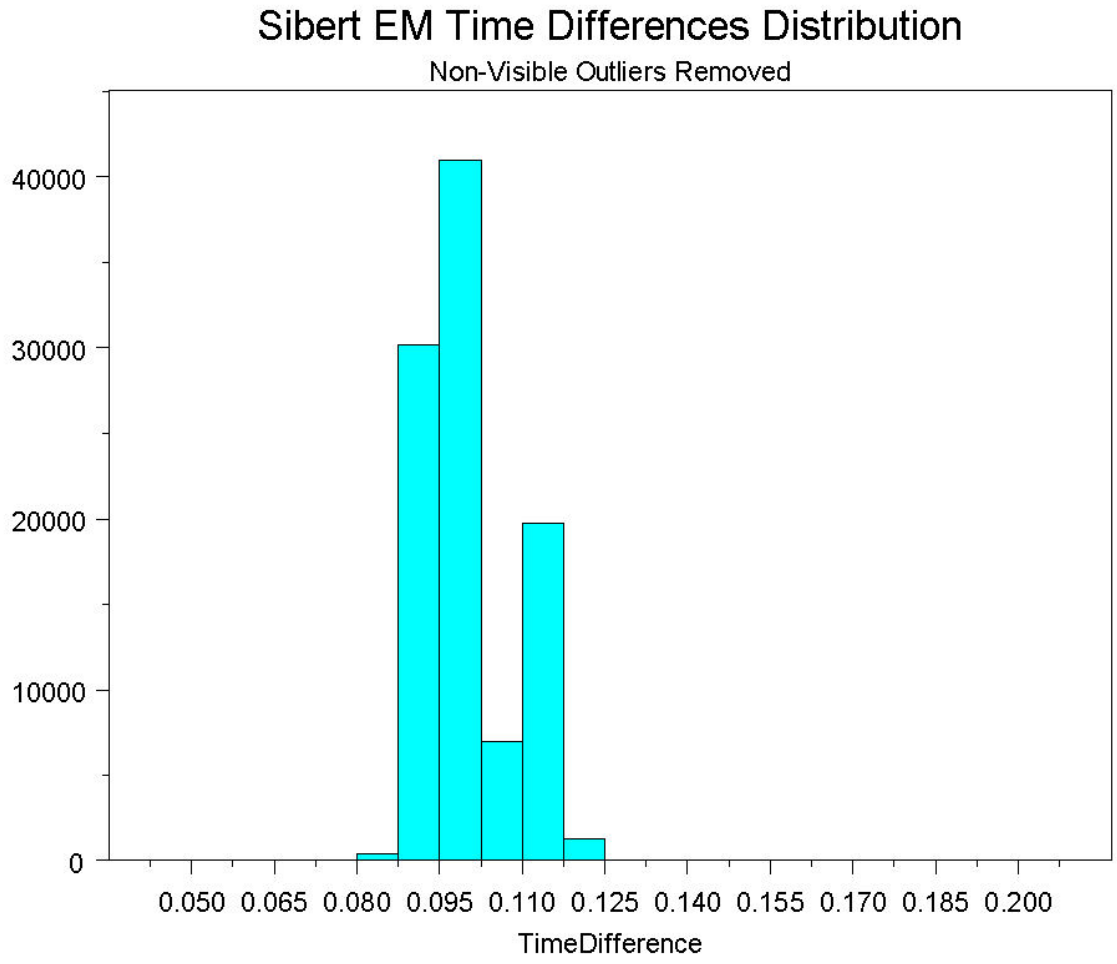


Figure 8. Time difference between 100000 randomly selected adjacent data points. Outliers excluded.



Both time and distance measurements are tightly distributed around a central tendency that is logical for this site.[16]

Accordingly, we concluded that the positioning of these data was sufficiently accurate and needed no further attention.

6.2.2 Leveling and Lag-Correction

The DGM for this track was delivered to us already leveled and lag-corrected. We did not make any changes in that regard.

[16] We note that the time differential is a bi-modal distribution. We have seen this same effect before at Warren A.F.B. This effect is a little odd and may be associated with the fact that data is taken in different direction passes with a slightly different speed for the different passes. In any event, given the tightness of the distribution, this issue did not concern us and was not further investigated.

6.2.3 Unexpected Data Issues

There were two issues in the Sibert data that required special handling for this project that significantly affected the data QAQC, the preprocessing, and the ellipse definition for each target. They were:

- Rut Noise; and
- Non-Target Anomalies.

These two data issues were important to the project because our discrimination approach requires that each target be defined by an ellipse that separates the target region from the background noise region (Section 2.1.3.2 and Section 6.3) Each of these two data issues made it difficult to *automatically* extract good quality ellipses for the program office targets or for non-program office targets.

Our solution to both of these data issues was, ultimately, to define the ellipses manually for each target and for each non-target anomaly. Having done that, the remainder of the LGP Discrimination Process went forward with little difficulty. However it took considerable effort to make that determination.

The next two sections will discuss each of these issues in depth and why they had to be handled with manual ellipse definition.

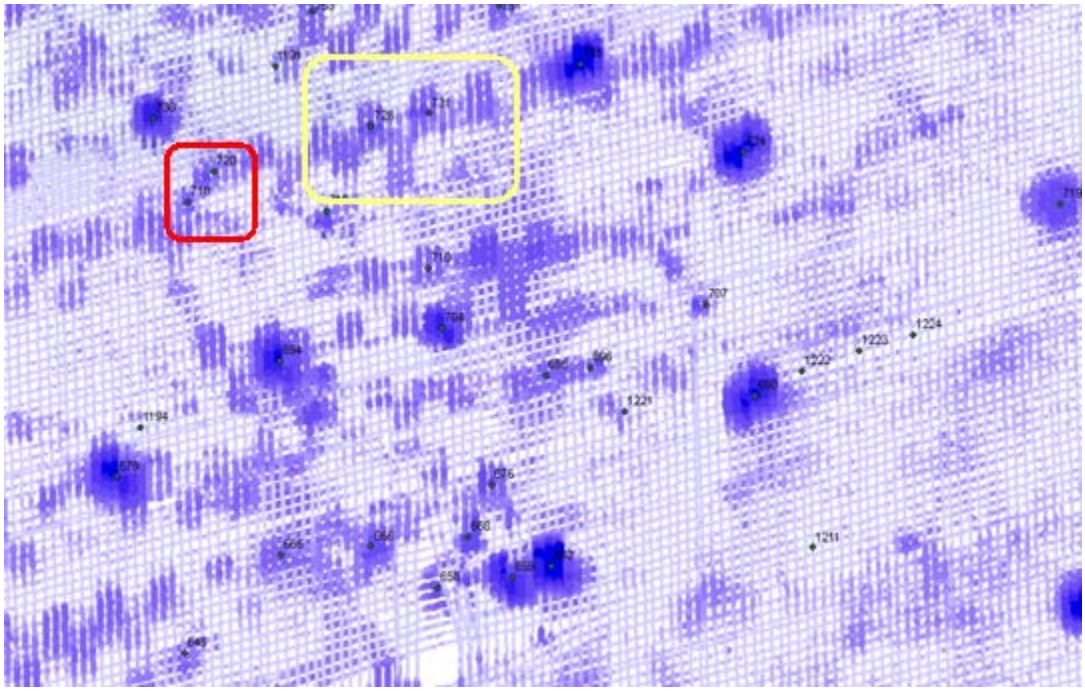
6.2.4 Rut-Noise

The rut-noise appeared as large regions of clearly above background noise signal that appeared to be oriented along straight lines. A plausible explanation for this is that the site contained regular, linear, ruts, such as might be made by vehicles driving over dirt in the same location repeatedly. That would, we expect, cause the bouncing of the array when it went over those ruts produced a signal, frequently non-trivial in amplitude.

Figure 9 shows a section of Channel 1 of the Camp Sibert DGM. This figure is not gridded—it shows individual DGM data points. The size of the individual data point is proportional to its millivolt value. More intensely colored regions, therefore, represent higher millivolt values. The small black dots with small numbers beside them are Program Office target picks. A few notes on this figure are appropriate:

1. The intensely blue regions in Figure 9 are clearly anomalous regions, all of which in this example were picked as EM Targets by the program office.
2. The rut-noise shows as medium-intensity purple regions in Figure 9. Note the linearity and mostly east-west orientation of the rut-noise. In fact, one can see the EM61MTADS array (groups of three north-south lines of data) as it hit an east-west rut from different directions. Different directions of movement (north or south) produced noise on different sides of the east-west rut (an example of that is highlighted in yellow) on Figure 9.
3. Targets often fell inside these regions of rut noise. In addition, many targets appear to have been picked because of high points in the rut-noise. A region in which that may have occurred for two targets is highlighted in red on Figure 9. And, the yellow highlighted region contains two additional targets.

Figure 9. Rut-noise in Camp Sibert EM61MTADS data



These regions of rut noise made it impossible to apply our typical preprocessing in a meaningful way. That process would normally be geared to normalizing the data so that good quality ellipses that define each target could be extracted automatically.

We spent a good deal of time attempting to adjust the process to behave properly; but our preprocessing assumes that anomalous regions may be distinguished from non-anomalous regions by reason of the fact that true anomalous regions are comprised of contiguous above-background noise data. Accordingly, we assume in preprocessing that the above-background noise region near a specified target fairly characterizes the specified target and the remaining signal (after removing the contiguous above-background noise) fairly characterizes the background noise and that the background noise is a reasonably stable distribution from target to target. The rut noise invalidated that assumption and we were unable to make adjustments to make the preprocessing algorithms work in the expected manner on these data. Ultimately, we elected to define the target ellipses manually and locate the non-target ellipses manually.

Table 7 illustrates the effect the rut-noise had on *non-target* portions of the signal in the vicinity of each target. This table was prepared using the manually defined ellipses for each target (as described below) and then removing all data points that fell within the designated target ellipses (see below) for both program office targets and non-target anomalous regions. In other words, we removed all data points in the region immediately around each program office target and immediately around each non-target anomalous region. What should be left is background noise.

To test whether the remaining data points comprised reasonably consistent background noise, we measured the 10% trimmed mean and standard deviation of the sum of

channels 1-3 from all out-of-target data points but that were less than eight meters from the identified location of program office targets. In other words, we retained data points in a donut around each target. The points in the donut hole, representing the target itself, were removed.

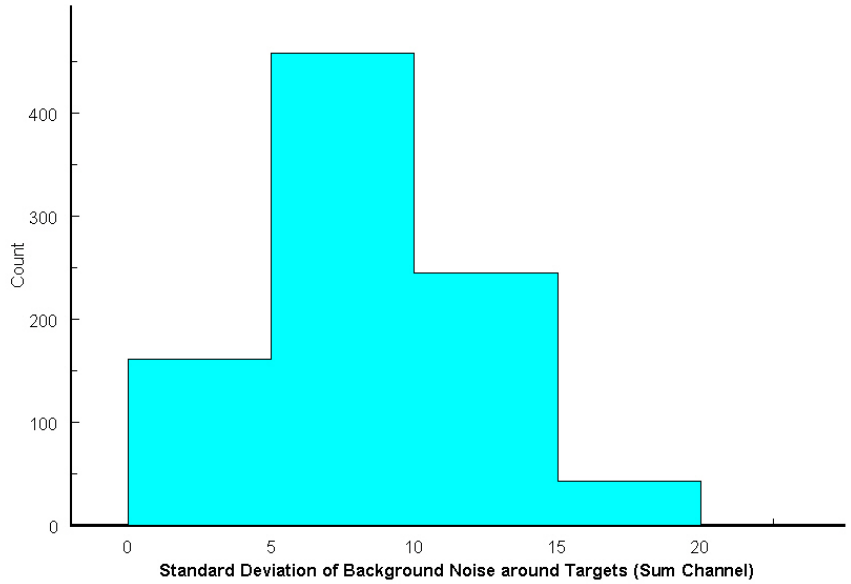
Ordinarily, having removed all above background noise targets data points from the signal, we would expect to see reasonably consistent and stable distribution of background noise. Table 7 illustrates how incorrect that expectation would be for this site.

Table 7. Statistics of the Background Noise across all Targets. Sum Channel (millivolts)

	Mean of Background Noise around all Targets	Standard Deviation of Background Noise around all Targets
Minimum	-12.189	3.78
Mean	-1.281	8.52
Median	-1.167	8.13
Maximum	7.587	19.68

The mean of the background noise amplitude ranged from -12.2 millivolts for Target 1253 to 7.6 millivolts for Target 880. That is an almost 20 millivolt variation in the mean background noise level amongst targets. Similarly, the standard deviation of the background noise ranged from 3.8 for Target 432 to 19.7 for Target 696. Another way to look at the instability of the distribution of the background noise is by way of Figure 10.

Figure 10. Histogram of standard deviation of background noise for sum channel



All targets whose background noise had a standard deviation greater than 10 had a 95% confidence range for amplitude of at least 40 millivolts (two times the standard deviation). Those with a standard deviation greater than 15 had a 95% confidence range

for background noise values of at least 60 millivolts. Either of these far exceeds acceptable background noise distributions for an EM61.

The spatial distribution of the background noise values was even more problematic. Figure 11 shows that spatial distribution for the standard deviation of the background noise across the entire site. Gray dots represent targets with below average standard deviation for the background noise surrounding the target. Red dots represent targets with above average standard deviation for the background noise. In addition, the dots become larger as the standard deviation becomes larger. The x axis is the zeroed X coordinate of the target. The y-axis is the zeroed Y coordinate of the target.

Figure 11. Spatial distribution of standard deviation of background noise

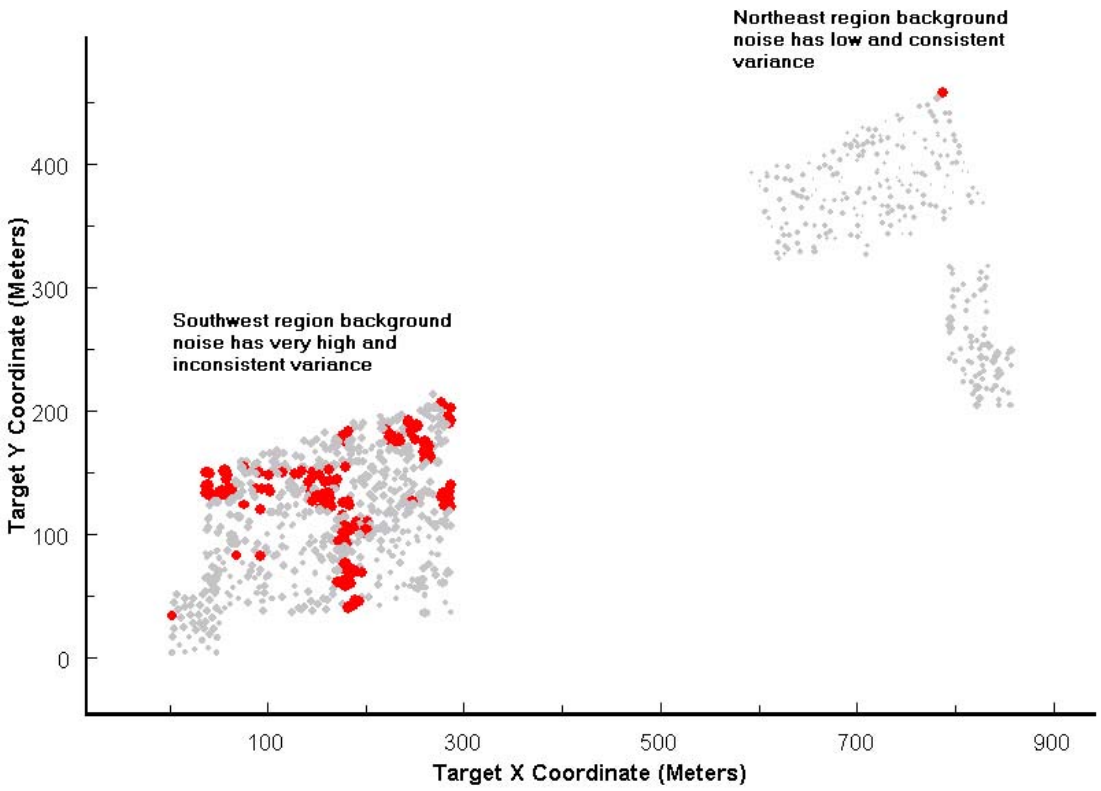
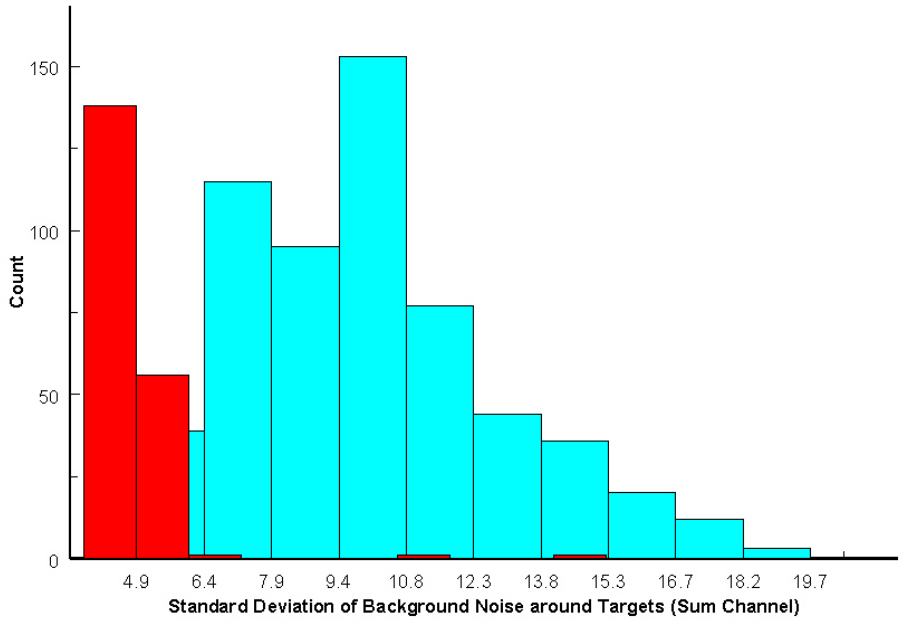


Figure 11 demonstrates rather clearly that the background noise standard deviation had much higher values in the southwest area than it did in the two northeastern areas. Note also the consistency of the values in the eastern regions relative to the consistency in the southwest region.

The degree of the disparity between northeast and southwest is shown in the superimposed histograms of the target-background-standard-deviations as between the large southwest area and the northeast area contained in Figure 12.

Figure 12. Superimposed histograms of standard deviations of target background noise in northeast area (red) vs southwest area (blue)



The northeast region data (red) is centered below five millivolts and is reasonably compact. It is within the range of what one would expect from good-quality EM61 data. The southwest region values (blue), on the other hand, have almost no overlap with the northeast region values and vary widely in a manner that suggests serious data problems in that region.

Like the above plots, visual inspection of the DGM also strongly suggested that the rut-noise was concentrated in the southwest regions (Figure 9 is in the southwest region, for example.). Thus, we concluded that the rut-noise was the probable source of the unstable background noise distributions and that the problem was widespread in the Southwest, and largest, area.

Visual inspection also suggested that the rut noise was mostly (not entirely) concentrated in the north-south of data (see Figure 9 and subsequent discussion).

6.2.5 Non-Target Anomalies

By “non-target anomaly,” we mean anomalous regions that were not designated as a target by the program office. We found that many anomalous regions (obvious targets) were not designated by the program office as targets.

All of our attribute extraction and preprocessing assumes that all anomalous regions have been identified as targets or have been excluded from the DGM. The program office did not designate a number of anomalous regions as targets. Initially, we thought to identify the non-target anomalies automatically. However, that because of the rut noise, that process produced poor results for such a substantial number of targets as to render it unusable.

Therefore, in order to identify non-target anomalies, we identified each such region manually from gridded Oasis Montaj data and marked its boundaries with a polygon. We then converted that polygon to the best-fitting ellipse using MSE as the error function and a downhill simplex optimizer in the same manner as we will describe later for the target ellipses. These ellipses were then treated as targets for the purpose of attribute extraction although they were not treated as targets for the purpose of discrimination.

6.2.6 Line Removal

Because the rut noise was so widespread in the Southwest region, we made only one significant alteration to the data we received by way of preprocessing. As noted above, visual examination strongly suggested that the rut noise was more pronounced in the North-South lines of data than it was the East-West lines of data. Accordingly, in that region, we removed the North-South lines of data.

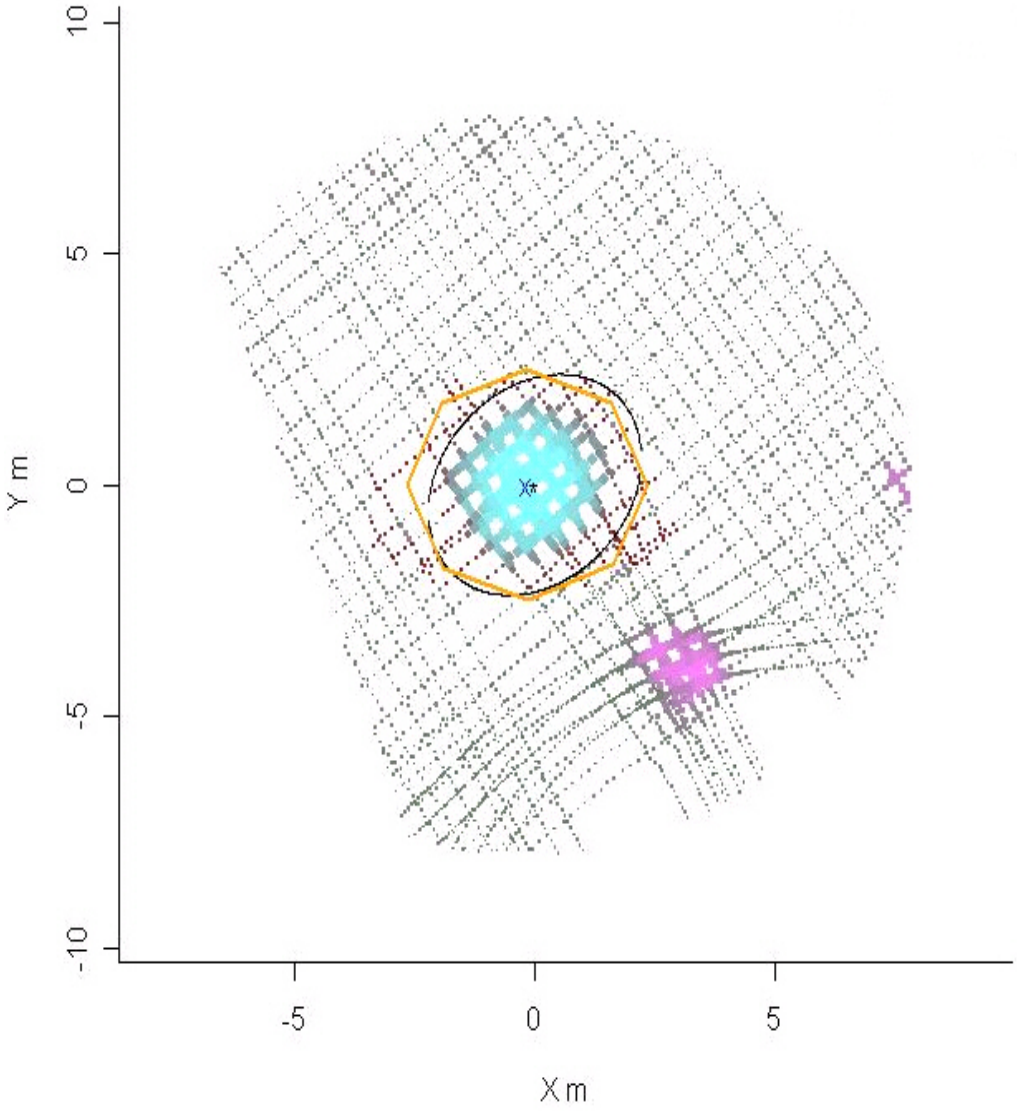
6.3 ELLIPSE DEFINITION

Each target (and each anomalous region that was not designated by the program office as a target) was defined by a single ellipse. The goal of ellipse definition is to optimally separate the background noise from the above noise signal. Ordinarily, we would generate the ellipses automatically. However, because the rut-noise created unstable and very different levels and variation in the background noise around the targets, we were not able to generate good ellipses automatically for a significant portion of the targets.

Accordingly, the ellipses were generated by two separate methods manual and automatic and the results were visually compared. The ellipse that best separated the target signal from the background noise was selected as the ellipse used for further analysis.

Figure 13 shows the results of both the automated and the manual steps in the ellipse definition process. The scale on both axes is meters away from the target pick center. Figure 13 superimposes this ellipse and this polygon over the DGM for Target 4 for channel 1. Each data point in the DGM is represented by a single point. The amplitude of the channel is represented by the size of point. Amplitude is marked with color also. In the target region, higher amplitude targets become bluer. Outside the target region, higher amplitude targets become more magenta.

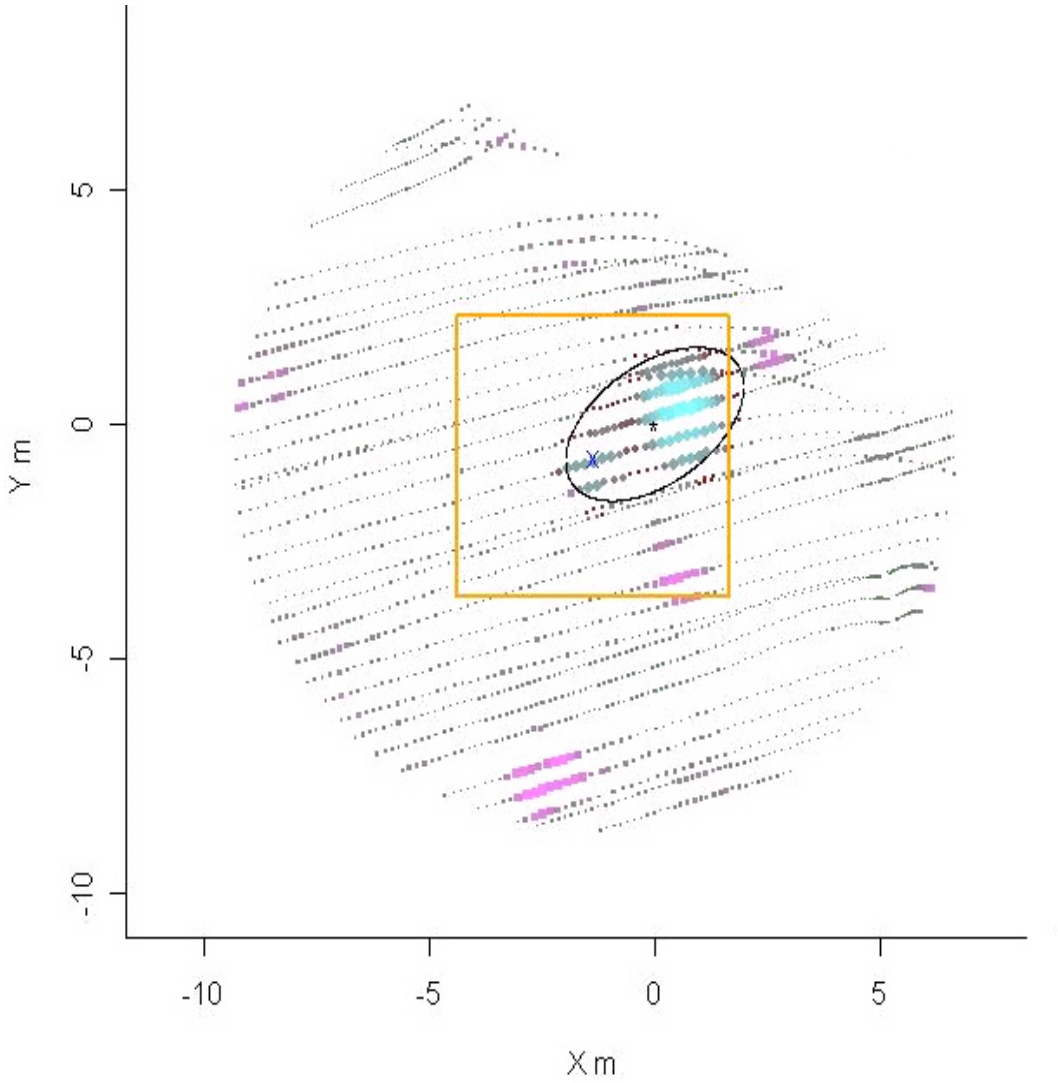
Figure 13. A successful definition of an ellipse and a polygon for Target 4. X and Y axes are zeroed on target pick location.



The black ellipse in Figure 13 is the automatically defined ellipse generated by our optimizer discussed below. The yellow polygon shows the manually defined polygon points connected by lines. This is a successful ellipse definition for both the polygon and the automatically generated ellipse. Either figure would be acceptable. We selected the ellipse as slightly better.

By way of contrast, Figure 14 shows the DGM around Target 1333, the automatically defined ellipse and the manually drawn polygon for the target. The colors and sizes of the objects are coded the same way as is Figure 13.

Figure 14. An unsuccessful attempt to define a polygon and an ellipse for Target 1333. X and Y axes are zeroed on target pick location.



In Figure 14, both the polygon and the ellipse definition failed. As we did not get a good target definition, this was one of our "cannot-analyze" targets.

The following sections on ellipse definition describe the process in more detail:

6.3.1 Manual Ellipse Definition

To define the target ellipses manually, the gridded Oasis Montaj data for each target and non-target anomaly was visually examined. Coordinates were selected so as to define a polygon that separated the anomalous points from the background noise point. The yellow polygon shown in Figure 13 is the polygon so defined for Target 4. We then converted the polygons into ellipses. The process for that was straightforward. We used a downhill simplex optimizer to find the ellipse that minimized the mean squared error between the vertices of the polygon and the ellipse. Table 8 shows sample output from that optimizer.

Table 8. Output of downhill simplex fit of ellipse to manually defined polygon--targets 1-34

TId	a	b	x	y	theta	MSE	Niters
1	1.6081193575...	1.6081193575...	622.25999569...	323.52893497...	0	7.30586475098172e-009	173
4	2.5000661659...	2.5000661659...	619.00999196...	327.52891096...	0	7.02835199612026e-010	161
5	1.306236024952	1.306236024952	627.11994427...	327.44895118...	0	7.07455760313092e-009	501
6	2.2980270448...	2.2980270448...	642.42003108...	329.50903423...	0	2.09350425485028e-008	149
7	2.6515216080...	2.6515216080...	655.90004130...	332.03915329...	0	4.53605242103154e-008	163
8	2.6250946057...	2.6250946057...	639.69009655...	335.67903259...	0	1.70700678480229e-008	501
9	1.5910228712...	1.5910228712...	619.78009402...	336.24906030...	0	1.26354762560073e-008	501
11	1.9100075349...	1.9100075349...	670.69000715...	339.05893140...	0	3.40735408771355e-009	187
12	2.6382852073...	2.6382852073...	678.98999165...	339.92890285...	0	3.36285577654516e-009	162
14	1.1811107229...	1.1811107229...	626.01999916...	342.20893761...	0	1.45424298267598e-009	501
15	2.6250946376...	2.6250946376...	631.09009655...	342.91903259...	0	1.70687102824165e-008	501
16	1.5910228685...	1.5910228685...	668.52009402...	343.34906030...	0	1.26354485848791e-008	501
18	2.6250945860...	2.6250945860...	639.92009655...	344.94903259...	0	1.70692046068321e-008	501
19	2.2980270448...	2.2980270448...	673.68003108...	345.97903423...	0	2.09350345891108e-008	149
20	2.6250859851...	2.6250859851...	624.66009962...	348.36902882...	0	1.76881068034039e-008	501
21	2.6515216080...	2.6515216080...	614.44004130...	348.55915329...	0	4.53605469636961e-008	163
22	2.7500503272...	2.7500503272...	667.35005480...	350.51876064...	0	9.95371170266749e-009	153
23	2.6382852073...	2.6382852073...	648.35999165...	351.71890285...	0	3.36285733619397e-009	162
24	1.3750423800...	1.3750423800...	628.04010597...	352.48894989...	0	4.30036521840649e-009	501
27	2.6515216080...	2.6515216080...	610.32004130...	356.59915329...	0	4.53605468394914e-008	163
31	1.6081193575...	1.6081193575...	685.55999569...	359.69893497...	0	7.3058645491169e-009	173
33	2.3365139784...	2.3365139784...	673.69996904...	361.64916886...	0	2.81257817450528e-008	501
34	1.1250035485...	1.1250035485...	606.29993354...	361.46894902...	0	4.20512744207817e-009	501
37	1.7848953635...	1.7848953635...	625.12004482...	363.65892727...	0	3.23076422156674e-009	499

The “TId” column is the program office’s target ID. The extracted parameters of the ellipse are:

- “a”: The semi-major axis of the ellipse in zeroed meters;
- “b”: The semi-minor axis of the ellipse in zeroed meters;
- “x”: The X coordinate of the ellipse in zeroed meters;
- “y”: The Y coordinate of the ellipse in zeroed meters;
- “theta”: The rotation of the ellipse in radians. Rotation is counterclockwise from an x-axis orientation.

The MSE column shows the mean squared error of the fit between the polygon vertices and the fit ellipse. These targets show a very close fit between the ellipse and the polygon.

6.3.2 Automated Ellipse Definition

Each EM61MTADS target was also identified by a parameterized ellipse that we extracted automatically. As discussed in Section 6.2.4, the automated process did not work well for many targets and worked well for others. The black ellipses shown in Figure 13 and Figure 14 show examples of the results of this process—one successful and one not-so successful.

We used the same parameterization for the automated ellipses as for the manually defined ellipses: (1) X coordinate of the center; (2) Y coordinate of the center; (3) Semi-major axis; (4) Semi-minor axis; and (5) Rotation in radians the x-axis.

The first step in the automated ellipse definition is to compute a z-score for the sum channel. The z-score is computed for all points in the eight meter circle surrounding the

center of a given target by first computing the background noise mean and trimmed standard deviation as described in the text accompanying Table 7. We then tagged points as above or not-above the background noise with a threshold of 1.75—in other words, about 1.75 standard deviations above the mean of the background noise. This is the step that the rut-noise affected strongly for many targets because the rut noise affected the standard deviation of the background noise in the vicinity of the target.

From the tagged data points, the ellipse for the target is then derived by Lipchitz Global Optimization from those tags for the data points in the eight meter circle around the target center. The objective function for the optimizer was the percentage of above-background noise data points in the ellipse. The result is an ellipse defined by the above five parameters that should, given good data, separate the above-background-noise points in the eight meter radius circle around the center of the target from the background noise that remain after removing the data points from all Targets from that circle.

We did not regard this portion of the project as successful. A substantial number of targets had automated ellipses that, on visual inspection, did not do a good job of defining the target (see Figure 14).

6.3.3 Selecting between the Manual Ellipse and the Automated Ellipse

The manual polygons and the automated ellipses were plotted against one another for each target as shown in Figure 13 and Figure 14. Similar plots were inspected for each target and the better of the two (the ellipse or the manually defined polygon) was picked and used for all further target identification.

6.3.4 Conclusion Regarding Ellipse Definition

In retrospect, we would have gone directly to manual ellipse definition and foregone the automated ellipse extraction and the attempted preprocessing that preceded it. The manual definition went far faster than expected and produced satisfactory results on almost all targets.

To our eye, the automated process produced slightly better results for well defined targets near stable background noise than did the manual process. However, when the target was near highly variable background noise areas, the automation failed and it was necessary to use the manually extracted ellipses on non-target anomalies to extract the automated target anomalies.

Nevertheless, it was the process of attempting to extract the elements of the automated ellipses that revealed the unstable distribution of the signal for background noise and permitted us to correct for it.

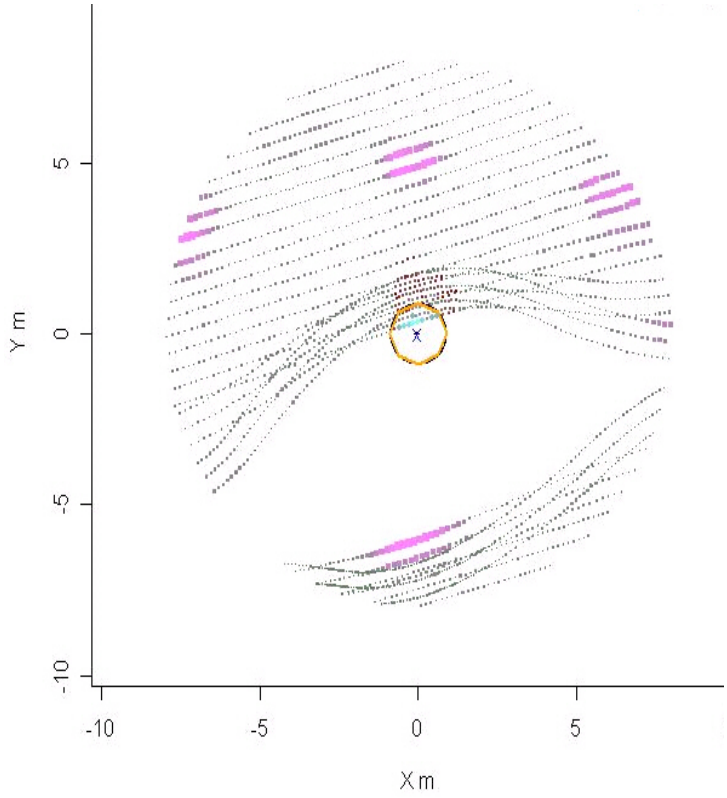
6.4 SELECTION OF CANNOT-ANALYZE TARGETS

Cannot-analyze targets were selected using six criteria, which are described in this section.

6.4.1 Insufficient Data

At the outset, we discarded targets where there was not enough data over the selected target to define an ellipse. This happened with respect to three targets in the southwest section. After the north-south lines were removed, there was not enough data remaining to generate a valid ellipse. Figure 15 shows Target 1290, an example of this situation.

Figure 15. Cannot-analyze target due to insufficient data (Target 1290). X and Y axes are zeroed on target pick location.

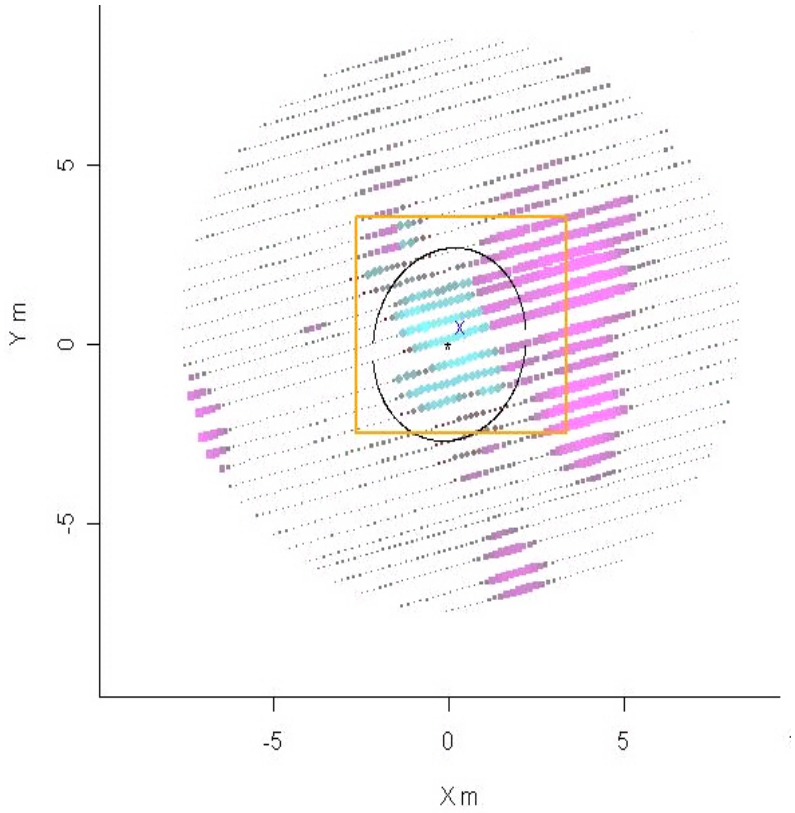


This criterion was applied after attribute extraction but before attribute reduction.

6.4.2 Ellipse Does Not Define a Target

For seven blind targets, we determined by visual inspection that the best ellipse produced by our ellipse definition process did not define anything that resembled a target. This occurred primarily in the southwest area. Figure 16 shows an example of this kind of target.

Figure 16. Cannot-analyze target because the ellipse does not define a coherent target (Target 1270)



We note that this cannot-analyze criterion was applied AFTER the amplitude discriminator (discussed below) was applied and only to targets that were above the stop-digging threshold of the amplitude discriminator.

6.4.3 Bad Ellipses

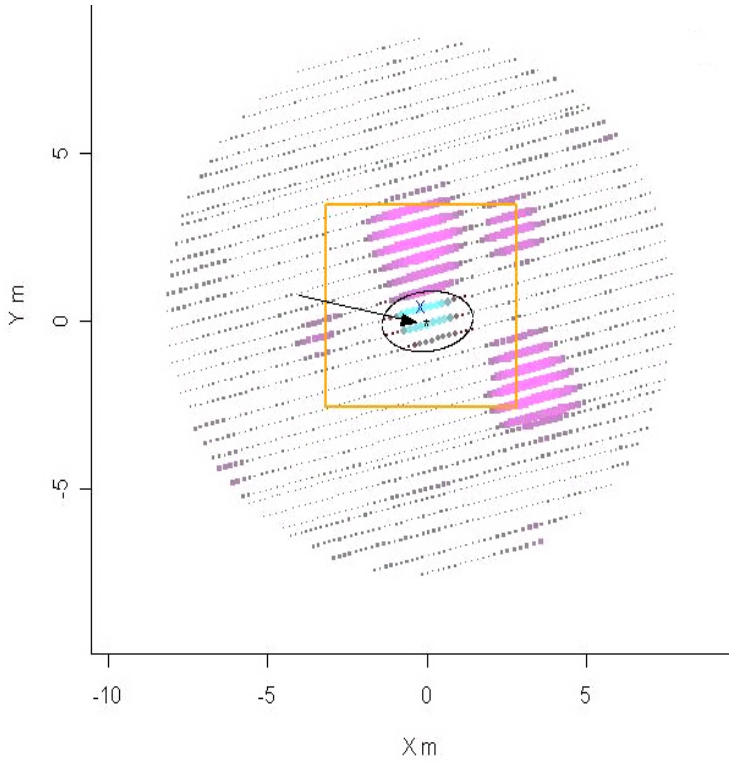
If the best ellipse (manual or automated) produced by our process was obviously wrong on visual inspection, we excluded the target. That happened for four blind targets. Figure 14, above is an example of that situation.

This criterion was applied AFTER the amplitude discriminator (discussed below) was applied and only to targets that were above the stop-digging threshold of the amplitude discriminator.

6.4.4 Overlap with Adjacent Target or with Adjacent Rut-Noise

Four blind targets posed possible overlap issues with either other targets or with rut-noise. Figure 17 shows an example of a probable overlapping target. The arrow points to the location of the program office pick. Figure 14 above would also have been excluded as cannot-analyze under this criterion.

Figure 17. Cannot-analyze target because of overlap. Arrow points to designated target location (Target 928). X and Y axes are zeroed on target pick location.



This criterion was applied AFTER the amplitude discriminator (discussed below) and only to targets that were above the stop-digging threshold of the amplitude discriminator.

6.4.5 Outlier Attribute on Important Attribute

Four blind targets had at least one attribute value that was an outlier on an attribute that was determined to be highly predictive of UXO in the attribute reduction process discussed below.

This criterion was applied after the attribute reduction process and before LGP modeling occurred. Examples will be shown in the attribute reduction discussion.

6.4.6 Insufficient Data Density in Attribute Space to Support a Do-Not-Dig Decision

Four blind targets were below the stop-digging threshold after LGP modeling but were designated as cannot-analyze because there was not sufficient data density in that region of attribute space to support the no-dig decision. Examples will be shown in the risk analysis section.

This criterion was applied after our risk analysis was complete.

6.4.7 Mistakes

We mistakenly assigned three targets to cannot-analyze.

6.5 ATTRIBUTE EXTRACTION

Attribute extraction is the process of converting the DGM in the vicinity of a picked target into meaningful statistics about the target. For this project, we extracted and used two types of attributes:

- Attributes that measure a statistic of the amplitude of the signal value of a single channel ("Amplitude Statistics); and
- Attributes that measure the ratio as between two different channels of Amplitude Statistics ("Ratio Statistics").

For the Amplitude Statistics, we measured a statistic for each channel plus the sum channel.

For the Ratio Statistics, we measured the channel ratios shown in Table 9:

Table 9. Measured Ratio Attributes

Numerator	Denominator
Channel 1	Channel 2
Channel 2	Channel 3
Channel 3	Top Coil
Top Coil	Sum Channel

Each of the statistics is measured in several different regions around the target location. The types of regions used are:

- The data points in the ellipse associated with the target ("entire ellipse region");
- The data points in ellipsoidal rings associated with the target ("ellipse ring region"); and
- The data in points circular rings associated with the target ("circle ring region").

The entire ellipse statistics are simple to describe: The statistic is measured for all the data points that are in the relevant channel and that are in the ellipse but that are not in nearby target ellipses.

Attribute extraction from ellipsoidal rings is a little more complex; but may be easily understood by viewing Figure 18.

Figure 18. a simple illustration of ellipsoidal rings for attribute extraction

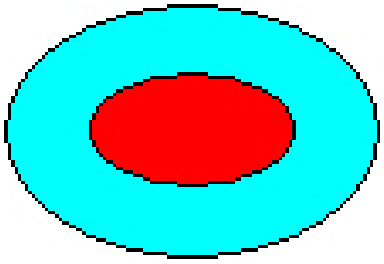


Figure 18 shows an ellipse (the dark outer boundary) associated with a selected target from which we wish to extract ellipsoidal ring attributes. This is of course the ellipse extracted under our ellipse extraction procedures. This figure shows two ellipsoidal rings defined by the ellipse, the inner ring (red) and the outer ring (cyan). The ellipsoidal ring attributes would be extracted separately from the data points located in each of these rings. That is, there would be a full attribute set of amplitude and ratio statistics for each of the two rings.

Circular rings are like the ellipsoidal rings. They are comprised of concentric circles centered on the target pick location. Each ring going out has a radius of 0.75 meters more than the next ring in. There would be a full attribute set of amplitude and ratio statistics extracted for each of the rings.

The statistics measured for every combination of region, Ratio Statistic and Amplitude Statistic were first, second and third moments.

6.6 ATTRIBUTE REDUCTION

The Attribute Extraction process described above produces hundreds of statistics for every target. The goal in attribute reduction is to reduce the number of attributes used in modeling to just a handful of highly relevant attributes that contain complementary information content about the modeling problem.

We used a collection of tools at different points in the modeling process to reduce attributes. The purpose of this section is to introduce the tools generally. We will describe how they were applied to particular problems in this project as we address those problems individually. The techniques include:

6.6.1 Numeric Input Binning

Binning numeric variables is a fundamental technique in machine learning. We use two sorts of binning in this project. Binning is the process of assigning numeric values to discrete categories:

Equal Frequency Binning. In equal frequency binning, a number of bins is specified and the numeric values are divided into that number of bins. This technique attempts to

assign the same number of numeric values to each bin. Sometimes that is not entirely possible because of tied numeric values.

Chi Squared Binning. Chi Squared Binning splits the numeric values into bins based on how well the splits do in minimizing the probability of Chi-squared statistic of the 2x2 contingency table formed by the split of UXO and Not-UXO on either side of the bin boundary. This is a recursive technique. It starts by finding the single split that has the lowest probability. If the probability is greater than a selected parameter, binning stops. If it is less, then each bin is split in the same manner. Splitting continues in each bin partition until the probability is greater than the set probability parameter.

6.6.2 Mutual Information

Mutual Information between an independent variable and the dependent variable (UXO) is usually one of the first measures we look at. Formally, the mutual information of two discrete random variables *X* and *Y* may be defined as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p_1(x)\, p_2(y)} \right)$$

We will refer to mutual information between a variable X and UXO as $I(UXO; X)$.

Typically, $I(UXO; X)$ is computed on a variable by variable basis and the results ranked. This gives a ranking of the variables that provide the most mutual information about the UXO/Not-UXO classification.

We compute *I* using discrete attributes and output. Accordingly, before any computation of *I*, it is necessary to bin the attributes first.

6.6.3 Maximum Relevance Minimum Redundancy

Maximum Relevance Minimum Redundancy methods ("MRMR") locate attribute sets with the maximum amount of mutual information between the attribute set and the target output and simultaneously, the minimum amount of overlapping mutual information as between the individual attribute in the dataset.[17] In other words, MRMR does not look for just the best attributes measured by mutual information between the individual attributes and the target output. Such attributes are frequently highly correlated and contain very similar information about the target output. Having five such attributes adds little or nothing to our ability to solve the problem. Rather, MRMR attempts to construct the attribute set that collectively contains the most information about the target output.

The MRMR algorithm is a greedy best-first algorithm. That is, it searches the entire attribute set for the single attribute that best increases the Relevance/Redundancy objective function. That attribute is added to the attribute set and that decision is not reexamined. Then the MRMR algorithm searches for the next attribute that, when added to the existing selected attribute set best maximizes the objective function. The size of the

[17] Hanchuan Peng, Fuhui Long, and Chris Ding, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, pp.1226-1238, 2005.

data set (*n*) is passed to MRMR as a parameter and the algorithm returns the *n* best attributes using the MRMR criterion.

We compute MRMR attribute sets using discrete attributes and output. Accordingly, before any computation of MRMR, it is necessary to bin the attributes first.

6.6.4 Correlation Based Feature Selection

Correlation-Based Feature Selection ("CFS") is very similar to MRMR. Its goal is to derive attribute sets that, collectively, do a good job of predicting the target output.[18] The differences are that CFS uses correlation coefficients instead of *I* as the measure of the predictive power of the attribute set and of the overlapping information included amongst the selected attributes. The advantage of CFS over MRMR is that it is not necessary to bin the attributes. The disadvantage is that CFS is not as good as MRMR at detecting non-linear relationships between attributes and the target output (UXO) and as between attributes selected for an attribute set.

We use CFS with a semi-greedy search algorithm. The algorithm adds the attribute that causes the largest gain in its objective function. However, unlike a purely greedy algorithm, our CFS algorithm is permitted to backtrack, that is, eliminate up to *n* of the most recently added attributes and start climbing from that spot. Obviously, if *n* is equal to the number of candidate attributes, then this is an exhaustive search algorithm, attempting all combinations of attributes.

6.6.5 Decision Trees

We use two forms of decision trees in variable reduction.

The first is the J48 single decision tree algorithm. It is an extension of the classic C4.5 decision tree algorithm.[19] J48 builds decision trees from a set of labeled training data using the concept of information entropy. It uses the fact that each attribute of the data can be used to make a decision by splitting the data into smaller subsets. The J48 algorithm may be summarized as follows:

> "J48 examines the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. To make the decision, the attribute with the highest normalized information gain is used. Then the algorithm recurs on the smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then a leaf node is created in the decision tree telling to choose that class. But it can also happen that none of the features give any information gain. In this case J48 creates a decision node higher up in the tree using the expected value of the class."[20]

---

[18] Hall M.A. Correlation-based Feature Selection for Machine Learning. Ph.D dissertation. Dept. of Computer Science, Waikato University, 1998.

[19] Ross Quinlan (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

[20] http://www.opentox.org/dev/documentation/components/j48

We use J48 as an alternative way to pick out attribute sets from MRMR and CFS. J48 is stronger at picking out interactions amongst attributes than is either MRMR or CFS.

Random Forests™ is a trademark of Leo Breiman. Random Forests is an ensemble decision tree algorithm that is reasonably fast and is does a good job of building preliminary models. We use Random Forests to assess the probable predictive result of a particular attribute set and also use its variable importance rankings as an attribute excluder. Random Forests is not particularly effective as an attribute includer.

6.6.6 Discipulus™ Input Impacts

After a project is finished, our core Discipulus Linear Genetic Programming software produces an "Input Impacts" report for that project. That report reports, for each attribute (input), what percentage of the best scoring evolved programs contained that attribute. It also measures how much each attribute contributes on average to the fitness of each of the thirty best evolved programs. We use these measures as attribute excluders.

6.7 PRELIMINARY ATTRIBUTE ANALYSIS

Our initial analysis of the attributes produced several important conclusions about how to model these data. It was comprised of two steps: minimal attribute reduction and graphic analysis of the best two attributes.

6.7.1 Preliminary Attribute Analysis—Attribute Reduction

We started by using Chi Squared binning on the attributes using a 0.99 confidence level for the splits. For this, we used only the training data.

Next, we submitted the binned data and the groundtruth labels to the MRMR algorithm selecting the attribute set consisting of the best ten attributes.

We then examined the best ten attributes and selected the two attributes that had the highest level of mutual information about the groundtruth labels. The two selected attributes had mutual information with the labels greater than 0.34. All other attributes had mutual information with the groundtruth labels of less than 0.15.

The attributes may be described as follows:

- V1: The first moment of the ratio of Channel 1 to Channel 2 in the outer ring of the ellipse;
- V2: The first moment of the ratio of Channel 3 to the Top Coil in the second circular ring out from the center of the target.

6.7.2 Preliminary Attribute Analysis—Results

Figure 19 shows both the training and blind data plotted in the V1, V2 attribute space. UXO are red, Not-UXO are green. The blind data is shown with small brown dots. We will use this format a good deal in the remainder of this report. For resolution purposes, extreme outliers are not shown.

Figure 19. Best preliminary attributes. Training and blind data with UXO and not-UXO clusters marked.

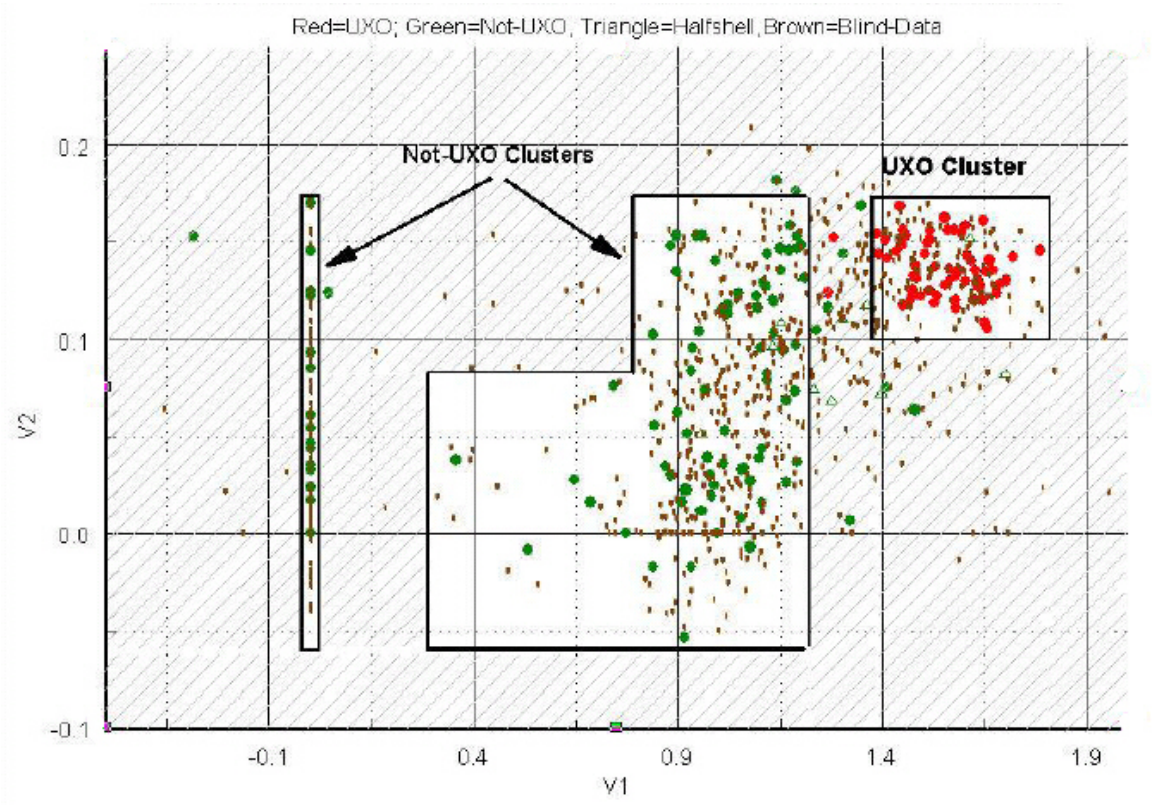


Figure 19 contains three distinct clusters of training data, defined by these two best features alone.

1. Cluster 1. The leftmost cluster contains only Not-UXO in the training data. It encodes the rule, "if either the major or minor axis of the ellipse < 0.75 meters, then the target is Not-UXO." This rule is a byproduct of our default value for an ellipse ring statistic where the total ellipse is too small to contain multiple rings.
2. Cluster 2. The center cluster also contains only Not-UXO in the training data.
3. Cluster 3. The rightmost cluster contains only UXO in the training data.

The three clusters at first glance do a very nice job of distinguishing UXO from Not-UXO.

Viewed another way, however, this is a not a particularly good attribute space for modeling.

1. The three cluster boxes define where we judge the training data appears to be dense enough to model. Outside the cluster boxes, 186 blind targets would have to be assigned as cannot-analyze.
2. The large number of blind data points that are well outside the regions containing training points strongly suggest a mismatch between the multivariate distribution of the training and blind data (on these attributes). In fact, 13.8% of the training data is outside the three cluster boxes but 25.4% of the blind data is outside the

three cluster boxes. When we analyzed the counts out of the boxes and total for training and blind data in a 2x2 contingency table, chi squared is 7.03 and the probability of that chi-squared is 0.008. The difference between the training and blind data percentages out of the cluster boxes is highly statistically significant.

3. The decision boundary around the UXO Cluster box is poorly defined by the training data. There appears to be a greater density of blind data in that region than training data.

We then examined the DGM for some of the targets in the outlier and decision boundary regions of Figure 19. We quickly concluded that the bulk of them comprised targets that were apparently selected because of rut-noise or where a small target DGM signature was intermingled with rut-noise.

What this means is that the mismatch between training and blind data was primarily on low-amplitude (meaning low-signal-value targets) and that the mismatch occurred on Ratio Attributes.

Accordingly, we determined that these data were best approached in two steps:

1. First, discriminate UXO from Not-UXO using only Amplitude Attributes. Determine which low-amplitude targets can be safely characterized as high-confidence not-MEC. Remove them from further analysis. The approach is thus to filter the low-amplitude targets first.
2. Then, using the remaining higher-amplitude targets, discriminate UXO from Not-UXO using our core LGP algorithm.

This is consistent with a fundamental rule of good modeling: Wherever possible, decompose the problem into multiple, simpler problems.[21]

The next two sections describe how we decomposed the problem into two discrimination task and the results in detail for each of the two sub-problems.

6.8 MODEL DATA WITH A SIMPLE AMPLITUDE DISCRIMINATOR

We performed the following steps to discriminate UXO from not-UXO using only Amplitude Attributes.

6.8.1 Designate Cannot-Analyze Targets

The “Insufficient Data” targets were removed as “cannot-analyze.” See Section 6.4.1.

6.8.2 Extract Amplitude-Only Attributes

We filtered our attribute set so that only those attributes from the EM attribute set that directly measure signal values were included. So, for example, all Ratio Attributes were excluded because of their instability on low-amplitude targets noted above.

[21] Langley, P. (1996) *Elements of Machine Learning* Morgan Kaufmann, NY, NY.

6.8.3 Amplitude-Only Attribute Reduction

Our goal in building an amplitude-based feature filter was to build a single independent variable model, using only amplitude-based attributes. The independent variable should rank a large number of Not-UXO at one extreme or the other. Using this single independent variable, we can rank the targets in terms of likelihood each target is UXO and, using the methods outlined below, convert that ranking into a probability of UXO and into a probability that UXO remain on site for each rank.

To identify that single, independent variable, we measured the mutual information between the training target labels and the binned amplitude attributes. For binning we used chi-squared binning and the 99% confidence level for the bin splits for this process.

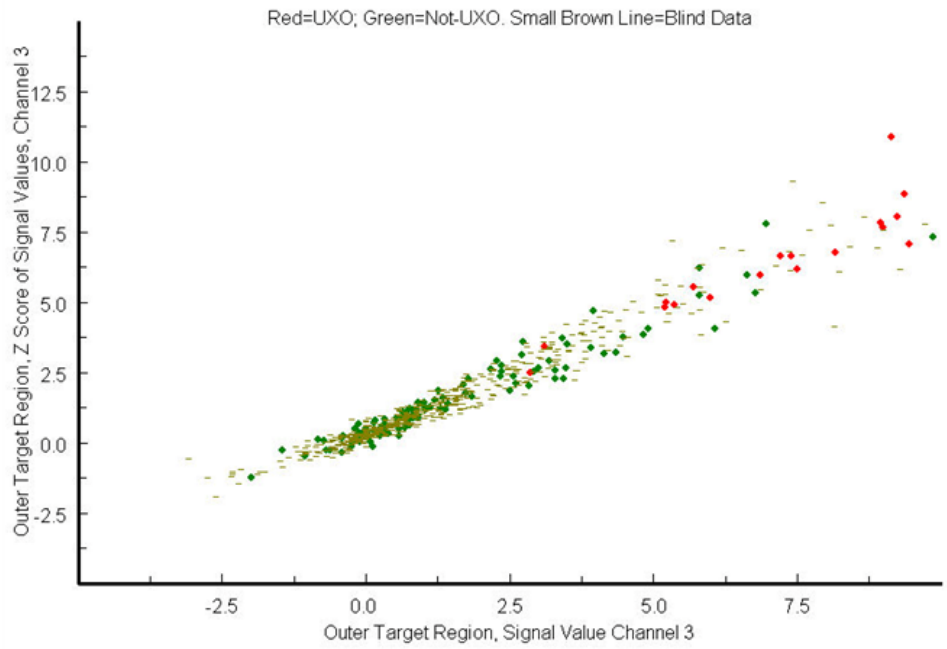
The two attributes with the highest level of mutual information with the training target labels were closely related and may be described as follows:

- AMP-V1: The Channel 3 (final decay channel) signal value in the outer region of the target ellipse. Mutual Information with training labels = 0.575.
- AMP-V2: The Channel 3 (final decay channel) signal value in the outer region of the target ellipse converted into a z-score relative to the distribution of the surrounding background noise. Mutual Information with training labels = 0.604.

Those two were selected as the basis for the amplitude-based discriminator.

Selected amplitude discriminator features on training and blind EM61MTADS data (close-up). Figure 20 shows the two best amplitude features and how well they discriminate the low-amplitude Not-UXO from UXO. AMP-V1 is shown on the X-axis and AMP-V2 is shown on the Y-axis.

Figure 20. Selected amplitude discriminator features on training and blind EM61MTADS data (close-up). X-axis shows AMP-V1 feature. Y-axis shows AMP-V2 feature.



On these two attributes, we have good discrimination for low signal value targets. For example, the lowest ranked UXO is at approximately 2.7 on AMP-V1. And, the distribution of the blind data (the small brown dots) matches the training data quite nicely.

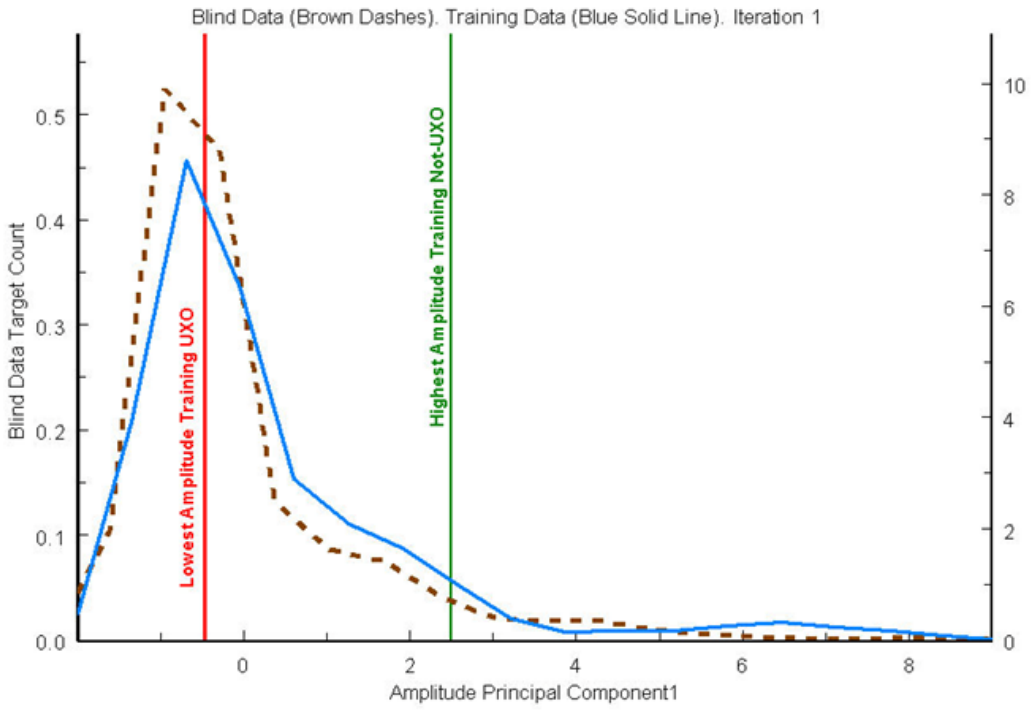
From only two features, it is quite simple to extract a single feature using principal components analysis. Effectively, the first principal component on these data projects each target onto the best regression line fitting the data, which is exactly what we want.

So we performed that principal component analysis and used the first principal component. The principal component used ("Amplitude Principal Component 1") may be described as follows:

- AMP-V1 is normalized with a mean of 6.69 and a standard deviation of 11.82.
- AMP-V2 is normalized with a mean of 6.11 and a standard deviation 10.26.
- Amplitude Principal Component 1 is 0.71 * Normalized AMP-V1 + 0.71 * Normalized AMP-V2.

As QAQC, we analyzed the distribution of the training and blind data as a function of Amplitude Principal Component 1. That analysis is shown in Figure 21.

Figure 21. Density of Amplitude Principal Component 1 on training and blind data

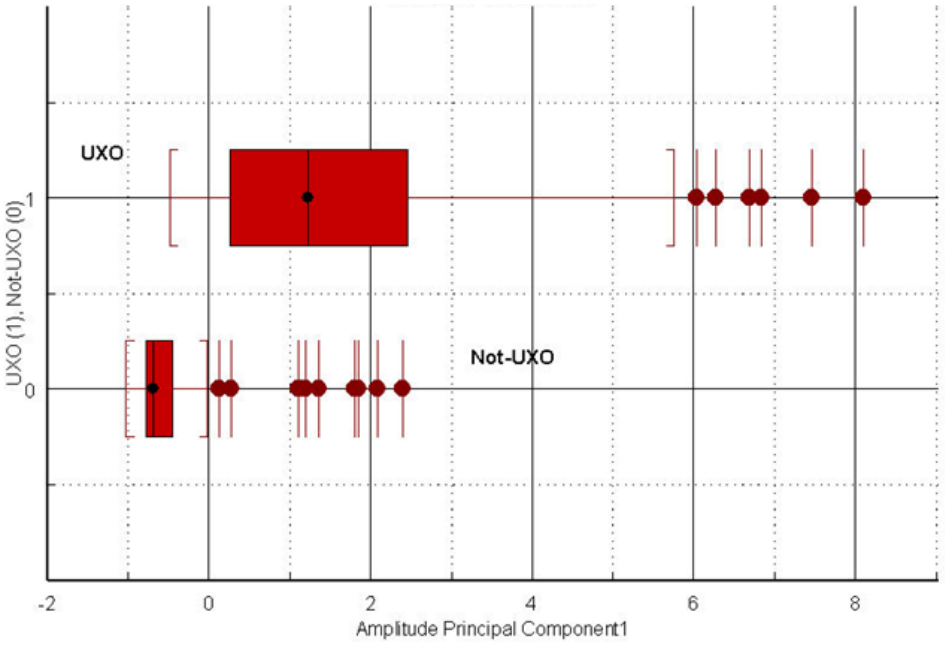


The match between the density of the training and blind data is quite close.

At this point we have reduced the amplitude attributes to a single attribute (Amplitude Principal Component 1), which, by itself provides a ranking. That is, the higher the value

of Amplitude Principal Component 1, the more likely an item is to be UXO. This is demonstrated in Figure 22.

Figure 22. Distribution of UXO and not-UXO on Amplitude Principal Component 1 training data. Comparative box and whiskers chart.



It is apparent that the great bulk of the UXO are concentrated between Amplitude Principal Component 1 values of 0.2 and 2.15. The UXO with the lowest Amplitude Principal Component 1 value is -0.47. On the other hand, the vast bulk of the Not-UXO is concentrated between -0.5 and -0.875.

The separation between classes is sufficient that Amplitude Principal Component 1 identifies a bin of Not-UXO that is highly statistically significant. The counts of UXO and Not-UXO above and below the split point are shown in Table 10.

Table 10. Two-by-two contingency table for best split on Amplitude Principal Component 1 on EM-only-track

	Below Split	Above Split
UXO	0	59
Not-UXO	84	32

The Chi Square statistic with Yates Continuity Correction for this table is 79.29 with one degree of freedom. The probability of that Chi Square is less than 0.0001.

6.8.4 Assigning Targets to High-Confidence Not-UXO Based on Amplitude Discriminator

To assign ranked targets to High-Probability Not-UXO, we applied our residual risk analysis approach to determine where, in the rankings provided by Amplitude Principal Component 1, we could safely say that none of the remaining items beyond that rank were likely to be UXO.

In this project, we are performing risk analysis at the 95% confidence level. As the amplitude discriminator adds an additional risk analysis step to what we had anticipated, we apply the Bonferroni correction to the confidence level used. Accordingly, we used a 97.5% confidence level in this and our risk analysis on the second modeling step described below. Together using 97.5% confidence level on the two steps will produce a Bonferroni corrected 95% confidence prediction of high-confidence Not-UXO.[22]

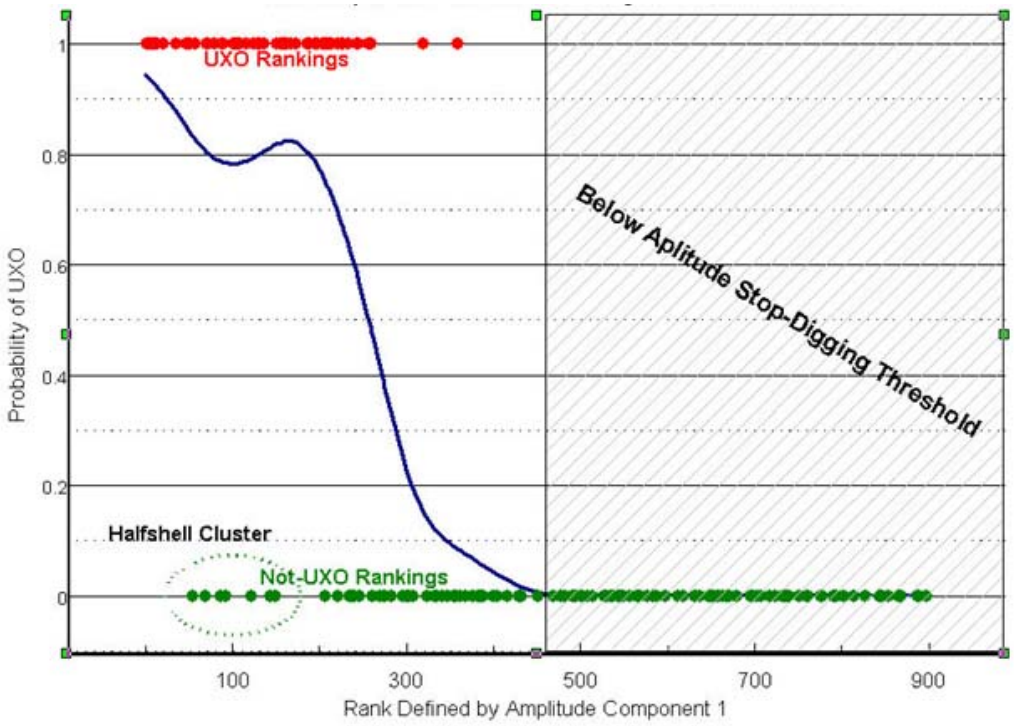
To perform this risk analysis, we converted Amplitude Principal Component 1 into a ranking across the training and blind data and used the ranking as our independent variable. (Rank 1 was the most likely to be UXO and higher ranks were less likely.)

We then modeled the falling probability of UXO as a function of this ranking. We assessed four different regression approaches—logistic, power-law, exponential and kernel regression.

We discarded the first three and selected kernel regression. None of the first three provide a decent fit to the falling probability of UXO as a function of ranking on these data. Figure 23 shows why.

[22] "[T]he Bonferroni correction is a method used to address the problem of multiple comparisons. It is based on the idea that if an experimenter is testing *n* dependent or independent hypotheses on a set of data, then one way of maintaining the familywise error rate is to test each individual hypothesis at a statistical significance level of 1/*n* times what it would be if only one hypothesis were tested." *http://en.wikipedia.org/wiki/Bonferroni_correction.*

Figure 23. Probability of UXO as a function of Amplitude Principal Component 1 Rank. Training Data



The red circles are the training UXO as ranked by the Amplitude Principal Component 1. The green circles are the training not-UXO ranked the same way. The blue line is the smoothed, local probability that a given rank is UXO. Note that at around ranking 180, the probably increases. The reason for this is the circled cluster of half-shells that the amplitude rankings find very early. The gap between that cluster and the remaining Not-UXO found causes the bump in the local probability value.

The result of this is that power-law, exponential and logistic fits are inappropriate as none of them will model the rise at ranking 180 well at all.

Accordingly, we used kernel regression to model the falling risk. It is an elegant technique that makes no assumptions about the form of the falling risk and requires only one numeric parameter, kernel width.[23]

To set this parameter, we used leave-one-out cross-validation on the training data. The kernel type used was a Gaussian kernel:

Equation 3: $$\mathrm{P}(UXO)_i = \sum_j e^{-\left(\frac{(x_i - x_j)^2}{2\alpha^2}\right)}$$

In Equation 3: (1) α represents the standard deviation (the width parameter) of the above Gaussian kernel; (2) x_i represents rank of the ith ranked blind data instance computed

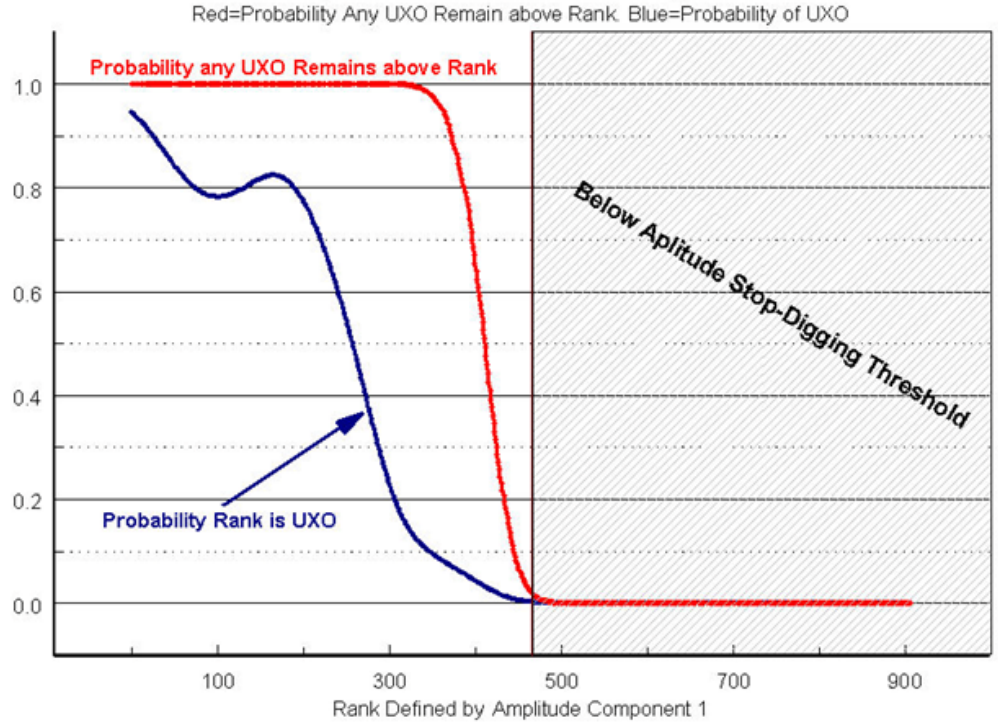
[23] Teknomo, Kardi (2007) *Kernel Regression*, http://people.revoledu.com/kardi/tutorial/regression/kernelregression

from Amplitude Principal Component 1 across all training and blind data points; and (3) x_j represents the rank of the jth ranked training data instance value of Amplitude Principal Component 1 across all training and blind data points.

We used a downhill simplex optimizer to minimize an objective function of minus two times the log-likelihood ("-2LL") of the regression results across the training data, assuming a Bernoulli distribution of errors, given a particular kernel width substituted into Equation 3. The kernel width parameter with the minimum value for -2LL on the held-out cross-validation set was 38.716. The unit is ranks generated by Amplitude Principal Component 1.

We then applied that derived kernel width parameter substituted into Equation 3 using, as the independent variable, the rankings of the blind data generated by Amplitude Principal Component 1. Figure 24 shows the results.

Figure 24. Kernel regression of probability of UXO as a function of Amplitude Principal Component 1 ranking. Blind data results.



The blue series in Figure 24 is the modeled probability of UXO as a function of the Amplitude Principal Component 1 derived rank.

The red series in Figure 24 is the cumulative probability that one-or-more UXO remain in any blind target ranked to the right of the plotted rank. It is computed for each rank using the "or of probabilities" approach described in Equation 2 in Section 2.1.6. Using that approach, the probability of one or more UXO remaining to the right of the measured ranking falls below 0.025 (97.5% confidence level) at ranking 463 (training and blind ranked together). This is equivalent to a determination that any target with an Amplitude Principal Component 1 value of less than or equal to -0.628 is high-probability Not-UXO.

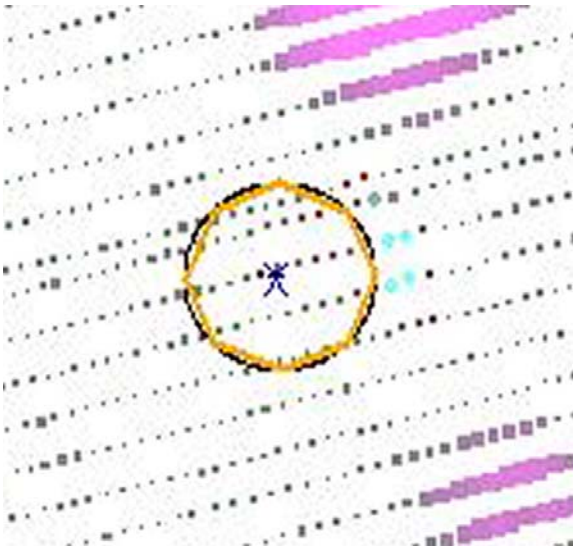
Given this cutoff, sixty-nine training targets fell into the high-probability Not-UXO region. Three hundred seventy-five blind targets fell into the high-probability Not-UXO region.

6.8.5 Effect of Amplitude Discriminator on Mismatch between Training and Blind Data from Preliminary Data Analysis

Recall that the reason we added an amplitude discriminator to our process in the first place was to address the mismatch between training and blind data distributions on some of the best features identified in Figure 19. The Amplitude Discriminator performed well in fixing this problem. We reached that conclusion for three reasons

First: Once targets were classified by the Amplitude Discriminator as high-probability Not-MEC, we spot-checked about 20 of the blind and training data outliers from Figure 19 against the amplitude discriminator rankings. Every one of the outliers was ranked as high-probability Not-UXO by the Amplitude Discriminator. An example of one of those outliers is shown in Figure 25.

Figure 25. Example of target designated as high probability not-UXO by amplitude discriminator (Target 840).



(For scaling reference, the defining ellipse for this Target 840 is a circle 1.5 M in diameter.)

Target 840 is not atypical of other such outliers from the boxed clusters in Figure 19. Most of these outliers came primarily from the Southwest Region, where rut noise was a substantial problem. Target 840 was apparently picked using the NS and EW lines of data. The NS lines of data were much more affected by the rut-noise than were the EW lines. So when we removed the NS lines of data to reduce rut noise, the above picture of the Target 840 DGM was all that was left and it produced implausible values on the Ratio Attributes.

Second: One side-benefit of the amplitude discriminator is that it found and identified another class of target (aside from the rut-noise targets) as high-confidence Not-MEC.

These may be described as relatively spiky and quickly decaying metallic signatures typical of smaller, thinner walled fragments near the surface.

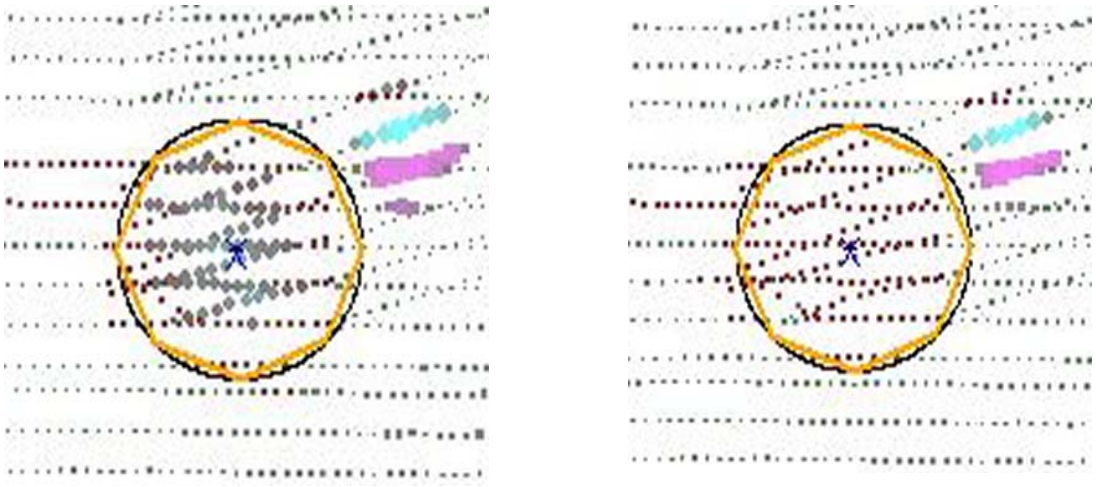
This result, in retrospect, is not surprising, for two reasons.

- To begin with the 4.2 inch mortars (and the half-shells) have relatively thick walls so they will decay more slowly than thinner walled objects. Accordingly, we would expect the final decay channel (used here) to be more affected by UXO and half-shells than by smaller, thinner-walled objects.
- Furthermore, 4.2 inch mortars will only show low amplitudes when they are deeply buried. The EM signatures of deeply buried objects tend to spread out and become less peaked.

Thus, deeply buried UXO on this site (the ones most prone to have low amplitude) should more strongly affect the last decay channel further away from the target center.

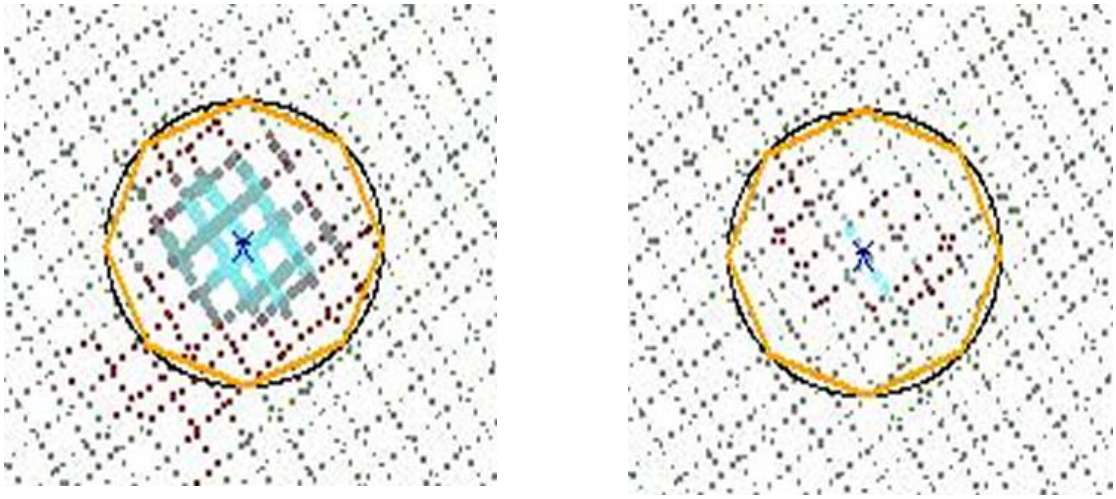
In fact, that is exactly what happens. The lowest ranked training UXO by the Amplitude Discriminator is Target 2014, shown in Figure 26. (Target 2014 is also the deepest of the training UXO and least favorably situated for detection.) It is low and wide on both the first and last decay channels. Although it decays from the first to final decay channel, its thick walls and the spreading of the signal due to its depth makes the outer part of the defining ellipse stand out clearly from the background even in the last decay channel (in this figure, larger dots are higher signal values). For scale, the circle shown below are about 2.5 Meters in diameter.

Figure 26. Deep 4.2 inch mortar signature on first (left) and last (right) decay channels



By way of contrast, Figure 27 shows a fragment that was classified as high-confidence Not-UXO by the Amplitude Discriminator. The first decay channel shows a substantial signal, far higher in amplitude than Target 2014 highlighted above. But this item, by the final decay channel has decayed to almost no signal, and none at all in the outer part of the ellipse.

Figure 27. Target 37, frag. First decay channel (left) and last decay channel (right)



In conclusion, our process picked the final decay channel in the outer region of the ellipses as the most important amplitude-based attributes. The results shown above are consistent with the expected physics of these attributes.

Third: We started the amplitude discriminator process primarily because: (1) We were uncomfortable with the number of outliers in the blind data on the more important predictive attributes on the training data; and (2) The low-density of training data in the decision boundary and the high-density of blind data in the same region.

Accordingly, our analysis of the amplitude discriminator ends with analysis of these same issues after removing the low-amplitude, high-confidence not-MEC targets.

Figure 28 shows the attribute space of the two most important attributes (V1 and V2) *before* the application of the amplitude discriminator.[24] We note again the large number of outliers in the blind data (the brown dots) from the distribution of the training data (the red and green circles).

[24] Figure 28 is a wide view of the same data shown in Figure 19. Best preliminary attributes. Training and blind data with UXO and not-UXO clusters marked. The effect of showing the wide view is to show extreme outliers in these data not shown in Figure 19.

Figure 28. Attributes V1 and V2. Attribute space before amplitude discriminator

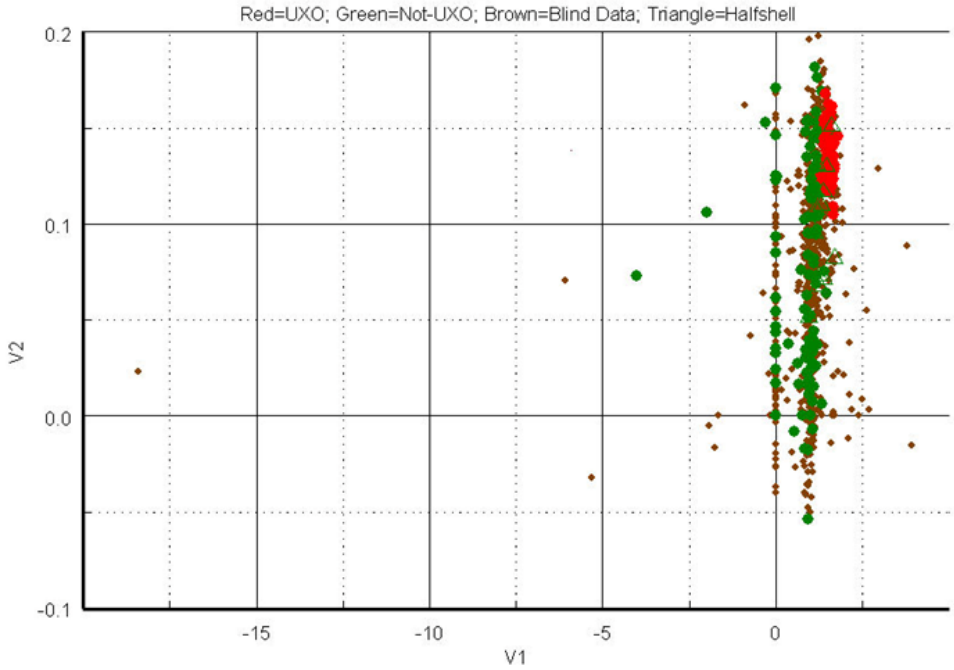
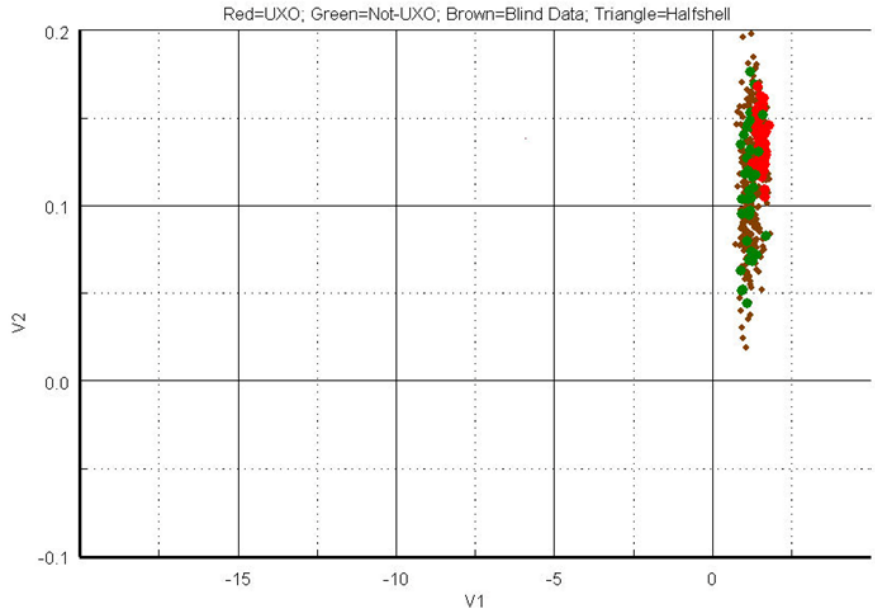


Figure 29 shows the same view as Figure 28 except that it is displayed after the application of the amplitude discriminator. After application of the amplitude discriminator, the extreme outliers shown in Figure 28 have disappeared and the input space appears much better condititoned.

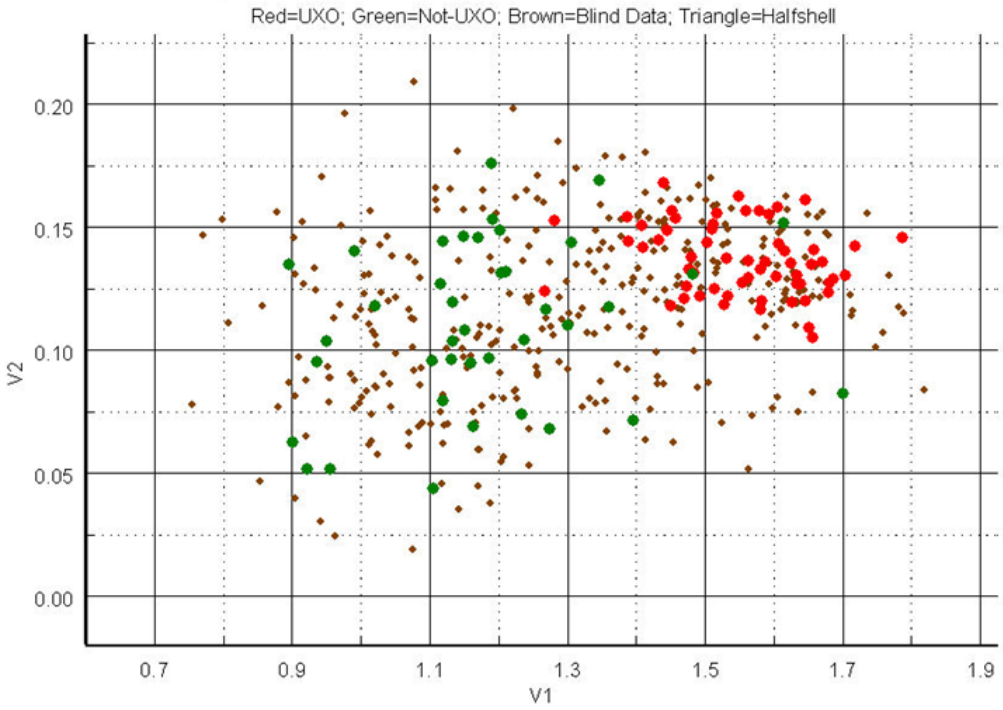
Figure 29. Attributes V1 and V2. Attribute space after amplitude discriminator (wide view)



Finally, Figure 30 shows a close-up of the same attribute space. They are reasonable well clustered and the distribution of the training and blind data matches nicely. It is particularly instructive to compare Figure 19 with Figure 30. The only difference between them is the

amplitude discriminator. The effect of the amplitude discriminator was to clean up the mismatch between training and blind data shown in Figure 19.

Figure 30. Close Up of Attributes V1 vs. V2. Attribute Space after amplitude discriminator (close-up view)



Fourth and Finally: A disproportionate percentage of the targets removed by the amplitude discriminator lay in the southwest area, which had the rut-noise. Only 59.8% of the targets *not removed* by the amplitude discriminator were from the southwest area. On the other hand, 74.4% of targets that were *removed* by the amplitude discriminator were from the southwest area. Table 11 shows the count.

Table 11. Count of targets above and below amplitude threshold in and out of southwest area.

	Below Amplitude Threshold	Above Amplitude Threshold
Southwest Area Count	262	216
Other Area Count	90	145

Recall that our hypothesized reason for the large numbers of outliers in attribute space *before* the amplitude discriminator was the uncontrolled rut-noise, primarily in the southwest area. The amplitude discriminator cleans up the outliers very nicely; and it does so by removing a disproportionate number of targets from the southwest area. This is consistent with our hypothesis about the source of the outliers.

Accordingly, in the remainder of the EM-only-track we will designate all items excluded by the Amplitude Discriminator as high-probability Not-UXO. Further modeling and discrimination will concentrate on the targets remaining after the application of the amplitude discriminator. We will refer to them as Above Amplitude targets or Higher Amplitude Targets.

6.9 MODELING UXO VS. NOT UXO WITH LGP FOR HIGHER AMPLITUDE TARGETS

This section describes the principal modeling task on this track. We applied the remaining steps of our process to this reduced data set of high amplitude targets as follows:

6.9.1 Target Exclusion

We removed from further consideration: (1) Cannot-analyze targets, as described in Section 6.4, except for the targets described in Sections 6.4.5 or 6.4.6; and (2) The high-confidence Not-MEC targets that were below the amplitude discriminator threshold. We were left with 98 Training and 339 Blind targets. The training data was then comprised of 59 UXO and 39 Not-UXO.

This section describes how we applied the remaining steps of our process to this reduced data set of high amplitude targets.

6.9.2 Attribute Extraction

We started with the same EM-only attribute set as we began our preliminary modeling pass.

6.9.3 Attribute Reduction

We applied the same multi-tool attribute reduction process described above.

The steps and were as follows:

We binned the data using chi-squared binning with a 99% confidence level..

We identified an initial set of attributes from the EM attribute set that had a high degree of mutual information with the UXO labels and that were optimally uncorrelated with each other using the MRMR algorithm. This set comprised 11 attributes.

Subsequently, we performed a preliminary ten-fold cross-validation LGP run using these initial 11 attributes. Three of the eleven attributes had a large and consistent impact on the solutions derived by LGP. Accordingly, those three attributes were selected as the starting point for our next step

We then re-binned the EM Attribute set using ten equal frequency bins. We used a semi-greedy best-first selection algorithm on the binned data with back-tracking set to 3 and searching set to both directions (that is, best-first will attempt to improve the data set by adding to the and by deleting attributes from the starting data set).. Each attribute set was evaluated using the Symmetric Uncertainty criterion.

We used, as a starting point for the best-first algorithm, the three attributes previously selected. We performed fifty-fold cross validation and recorded the percentage of folds in which each attributed appeared as part of the optimal input set. Seven attributes appeared in at least 50% of the folds and they were accepted for the next step. All three of the initial attributes were selected by this additional step.

We then performed an additional ten-fold cross-validation LGP run using these seven attributes. One of them had no impact at all on the LGP solutions and was discarded. The remaining six attributes comprised the basis of the training set for all further EM LGP runs in this iteration.

At this point, there is a degree of convergence as between different attribute selection procedures. Three of the six final attributes were the original three attributes selected by the MRMR algorithm above. Additionally, of those six attributes, four were identified by our subsequent LGP models as highly significant in discriminating UXO from Not-UXO on this site and two of possibly non-trivial importance. Table 12 shows the results of LGP's attribute impact analysis in its final ensemble model for Iteration 1.

6.9.4 Graphic Analysis of Best Attributes

The first four attributes selected by the above process, by themselves show good class separation between UXO and Not-UXO. In addition, when graphed, the distribution of the blind data matches the distribution of the training data reasonably well.

Figure 31 shows the distribution of the two variables that the above process identified as most important, V1.1 and V1.2. The training UXO are shown in Red. The training Not-UXO data are shown in Green. The blind data are represented by the small brown dots.

Figure 31. EM-only-track: Two most important attributes for LGP modeling. Training and blind data.

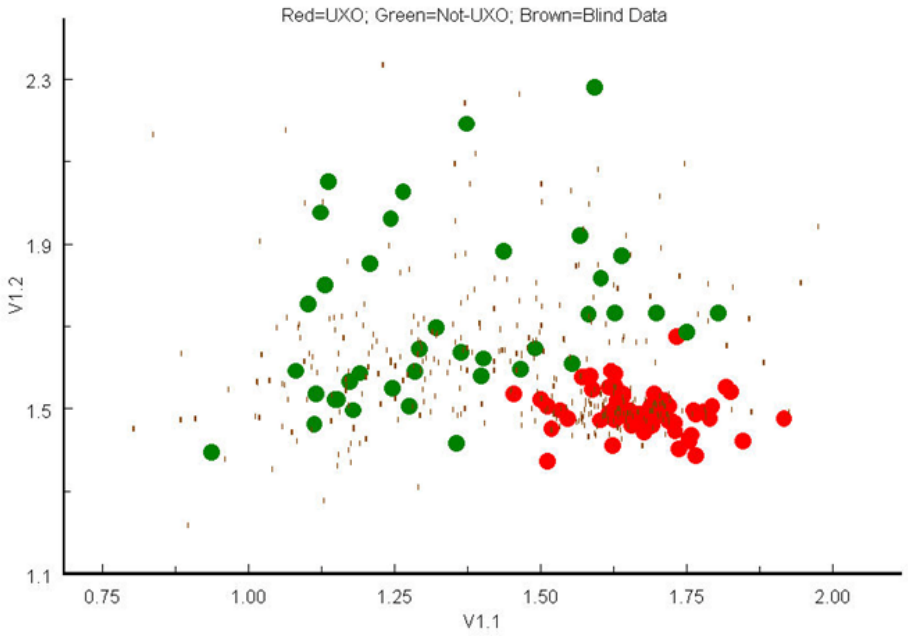
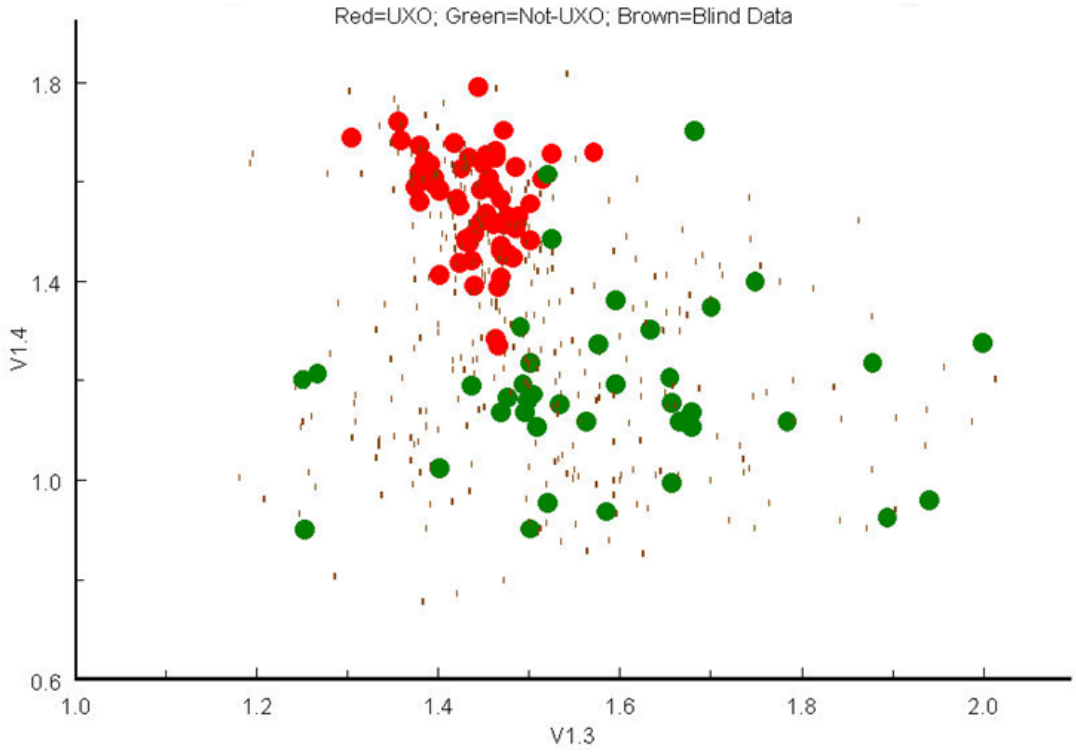


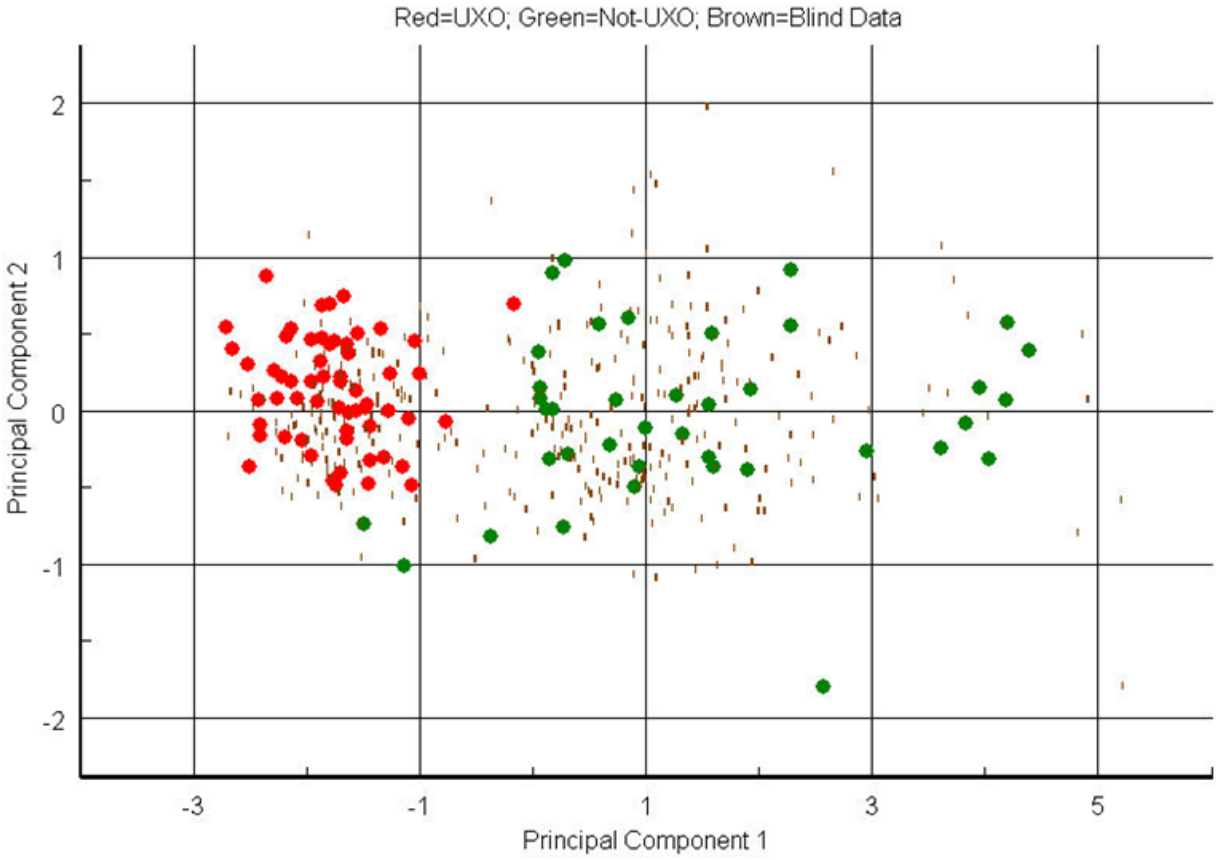
Figure 32 shows the distribution of the two variables that LGP identified as third and fourth most important, V1.3 and V1.4. The training UXO are shown in red. The training not-UXO are shown in green. The blind data are represented by the small brown dots.

Figure 32. EM-only-track: Third and fourth most important attributes in LGP modeling. Training and blind data.



Having selected the attributes, we then further reduced the dimensionality of the problem for visualization by the use of principal components. Figure 33 shows selected principal components of the six selected attributes. The class separation is almost perfect and we proceeded to build our final models on these data.

Figure 33. EM-only-track: Principal Component 1 vs. Principal Component 2 of six selected attributes for modeling



The class separation between UXO (red circles) and Not-UXO (green circles) is almost perfect. Further, the match between the training (red and green circles) and blind data (brown, small circles) is very good. Accordingly, we determined that this project was ready for final modeling with LGP using these six attributes.

6.9.5 LGP Modeling Procedures

This is a small data set. The biggest danger is overfitting to the training data and producing models that do not generalize well to the blind data. Dimensionality reduction was the first important step to preventing overfitting. Here are the additional steps we took to build models and minimize the danger of overfitting.

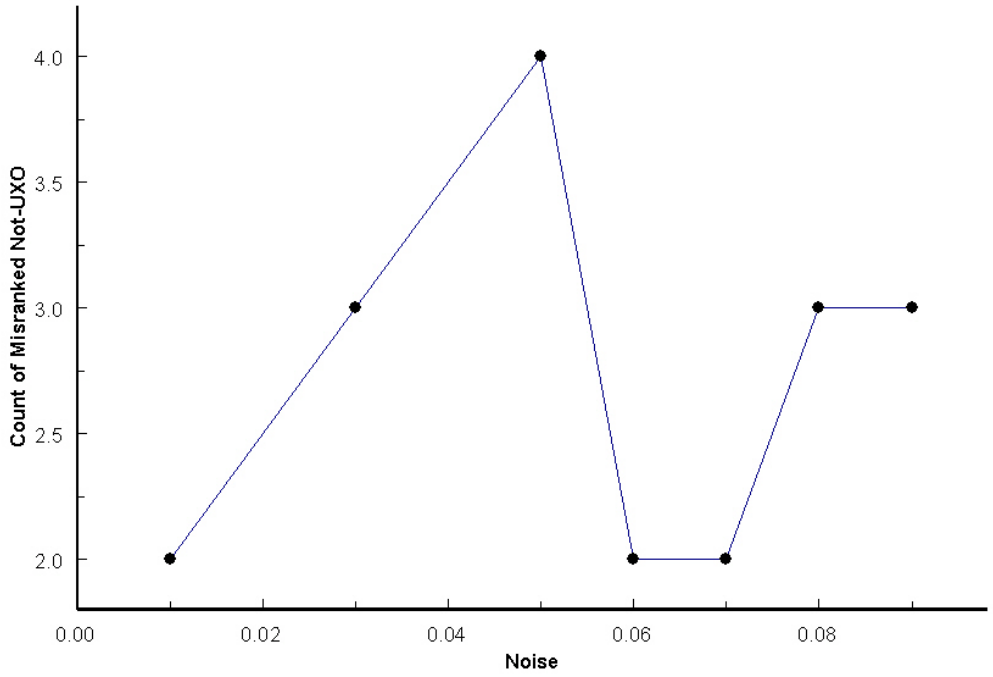
The most important modeling decision was that the data set was small enough that we should add noise to the attributes to prevent overfitting. We replicated each row in the training data 30 times and added a small amount of noise to each input, defined by a percentage –from 2% to 9%. Adding noise in inductive modeling is equivalent to Tikhonov Regularization and, if the correct noise level is selected, reduces overfitting.[25]

[25] Bishop, C. (1995) "Training with Noise is Equivalent to Tikhonov Regularization." *Neural Computation* 7 No. 1 (1995) 108-116.

We selected the noise parameter by performing ten-fold cross-validation projects at selected noise levels from 1% to 9%. Discipulus was set to its default parameters, except for the following: (1) Fitness function was set to "AUC;" (2) Each run in the project was terminated at 40 generations without improvement; and (3) The number of runs in each project was 20. At the end of each project/fold, we opened the program designated by Discipulus as the best program of the project and we repeatedly removed introns from that program until the best program ceased getting shorter. The best program with introns removed was selected as the program model for that fold. Its scores on the held-out data for that fold were stored. After all ten cross validation projects were completed, the stored scores were aggregated and targets with multiple scores were assigned a score equal to the average score for that target. This provided a single score for every target, which we interpret as a ranking.

Figure 34 shows the results of the 10 cross-validation projects in terms of how well they rank the held-out cross-validation data. It shows how many Not-UXO were ranked above the lowest UXO by noise level. Obviously, a lower value is better.

Figure 34. Count of misranked not-UXO by noise level using ten-fold cross validation.



The Area under the curve for the ROC curves generated for the various noise levels for these rankings ranged from 0.9784 (at 8% noise) to 0.9969 (at 7% noise). These are all excellent values and we would expect any of them to produce good models. We selected 6.5% as the noise level as it fell between the best two adjacent noise levels per the cross-validation runs. Accordingly, that noise parameter (6.5%) was selected for further modeling.

We then performed 30 bagging projects with Discipulus LGP using 6.5% noise, with the in-bag set-size equal to the size of the training input set. For Discipulus, we used the same parameters described above for the cross-validation runs.

At the end of each project/bag, we opened the program designated by Discipulus as the best program of the project, we repeatedly removed introns from that program until the best program ceased getting shorter. The best program with introns removed was selected as the program model for that bag. Its scores on the out-of-bag data for that fold were stored. After all thirty bagging projects were completed, the stored scores were aggregated and targets with multiple scores were assigned a score equal to the average score for that target. This provided a single score for every target, which we interpret as a ranking.

In addition, after each project/bag was completed, we stored the scores of the same program on the blind targets. This produces multiple scores for each target. The average score for each blind target was treated as the predictive ranking for that target.

At the end of this process, we had constructed an LGP ensemble predictor, comprised of 30 evolved programs from LGP, each of which had been trained on a different sample from the training data set. The outputs from those thirty programs was reduced to a single predictor for the training and blind targets.

6.9.6 LGP Modeling Results on Training Data

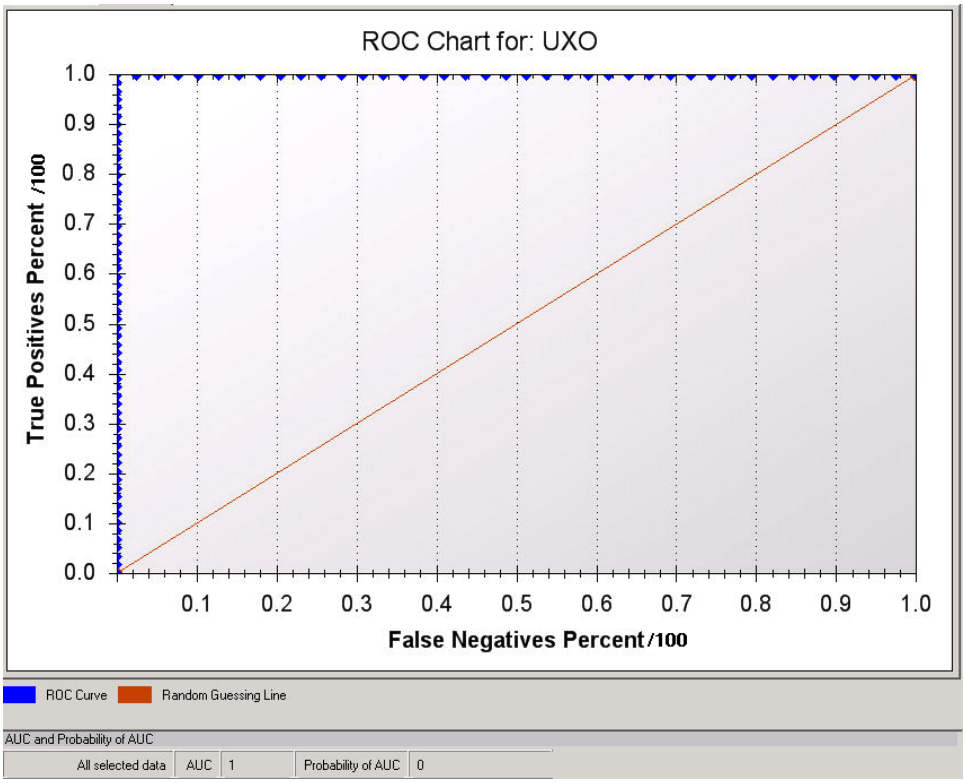
This section summarizes the performance of the LGP ensemble predictor on the training data.

The results of the ensemble predictor on the held-out training data are as follows:

1. All UXO in the training set are correctly ranked
2. All Not-UXO in the training set are correctly ranked

The AUC of the ROC curve is, therefore 1.0. This is perfect discrimination. Figure 35 shows the ROC curve for these predictions.

Figure 35. ROC chart for held-out training data for EM-only-track. LGP ensemble predictor ranking.



6.9.7 Attribute Importance

LGP produces an Input Impact report in each project. It describes the percentage of the best 30 programs of each project that contain each attribute in the project (frequency column). It also measures the average and maximum impact on fitness (over the best thirty programs) of each input (average and maximum columns). We output that Impact report from each of the 30 projects/bags in the ensemble predictor and then summarized the results across all bags by averaging the values.

In summary, LGP identified four attributes as the most important attributes in arriving at a perfect discrimination solution on this iteration. For ease of reference, we will refer to these attributes as V1.1 – V1.4 inclusive.[26]

Table 12 summarizes the impact of the six selected variables on the final EM-only ensemble predictor over the thirty bagging projects that created our final model.

[26] The remaining two attributes out of the attribute set had little influence on the solution. Removing them from the evolved solutions resulted in statistically insignificant changes in the generated ROC Chart. For example, removing V1.5 from the solutions changed the AUC of the best programs by only 0.003 on average.

Table 12. Variable importance analysis for EM-only-track

	IMPACTS			RANKS		
Input	Frequency	Maximum	Average	Frequency	Maximum	Average
V1.1	0.997	0.1139	0.0356	2	1	1
V1.2	0.973	0.0563	0.0132	2.2	2.7	2.6
V1.3	0.953	0.0462	0.0094	2.6	3.7	3.2
V1.4	0.53	0.0388	0.0116	4.6	3.8	3.9
V1.5	0.39	0.0094	0.0030	4.8	5	5.3
V1.6	0.347	0.0140	0.0049	5.4	4.8	5

There is a clear break in importance between V1.4 and V1.5, that boundary is marked with a darker line.

V1.1 through V1.4, the significant attributes on this EM-only-track, may be described as set forth below:

1. The first moment of the ratio of the top coil value to the sum channel value in the entire target ellipse;
2. The first moment of the ratio of channel 2 to channel 3 in the center ring of the ellipse;
3. The first moment of the ratio of channel 2 to channel 3 in the entire ellipse; and
4. The first moment of the ratio of the top coil to the sum channel in the center ring of the ellipse.

6.10 RISK ANALYSIS

This section describes the application of our risk analysis methodology to the LGP ensemble predictor described in the previous section for the EM-only-track.

In summary, we took the scores of the LGP ensemble predictor for both training and blind data for this step and combined them to produce a combined ranking across both data sets. In making that conversion from scores to ranks, a low LGP score was converted to a high ranking (that is, a low LGP score translates to a ranking that is less likely to be UXO). Then, we built a parameterized logistic regression model of the probability of UXO as a function of the combined rank, using that combined rank and the known groundtruth for the training data. Finally, we applied that parameterized model to the blind data and calculated the residual risk from the resulting probabilities for the blind targets

6.10.1 Risk Analysis Model Built on the Training Data

After assembling the combined ranks for this track, the next step in risk analysis was to build a probabilistic regression model of the UXO/Not-UXO groundtruth as a function of the rank across the training and blind data in this step. To build the model, we used the training data and associated groundtruth labels.

The four functional forms we considered for risk analysis were: exponential fit, power law fit, logistic fit and kernel regression. We immediately discarded exponential or power law fits to

model probability in this track. Both are monotonically decreasing functions with a continuously increasing first derivative. The perfect ranking on the training data in this track was better represented a step-like function.

Accordingly, the obvious functional form to use here was a logistic function derived using logistic regression. It is flat at both ends, that is, the first derivative of the function may be positive or negative at different points along the x-axis. This permits more step-like behavior.

Logistic regression, however, presents a numeric problem when confronted with a perfect ranking. Logistic regression uses optimization across all *i* training instances to determine α and β parameters of the following function:

Equation 4: $$\ln(P_{UXO_i} / (1 - P_{UXO_i})) = \alpha + \beta \cdot Rank_i$$

Unfortunately, the maximum-likelihood fit to a perfect ranking is when the β parameter approaches infinity, or a vertical line. As a result, the logistic regression optimizer we used pushes the solution toward an infinite slope and produces NAN's on a perfect ranking.

The solution to this numeric issue was based on the following observation. An imperfect ROC chart produces a more conservative risk assessment than a perfect ROC chart. That is, the slope of the logistic line will be less for an imperfect ROC chart, which will, as a result, assess the near-zero risk zone as being further down in the dig-list than a perfect ROC chart.

Accordingly, we determined, empirically, the minimum imperfection in our prioritized-dig-list rankings that did not produce numeric overflow in the logistic regression. We did so by reversing the label as between the top-ranked Not-UXO and the bottom-ranked UXO. We had to perform this step twice before we were able to derive a logistic solution that did not overflow.

The result is a logistic fit of probability of UXO as a function of the combined rank that somewhat underestimates the number of Not-MEC that may be safely left in the ground. Underestimating that risk is better than overestimating it, given the cost of a False-Negative. And that is the closest approximation of the declining probability of UXO on these data to the correct, but numerically impossible maximum likelihood solution.

The parameters of that fit are:

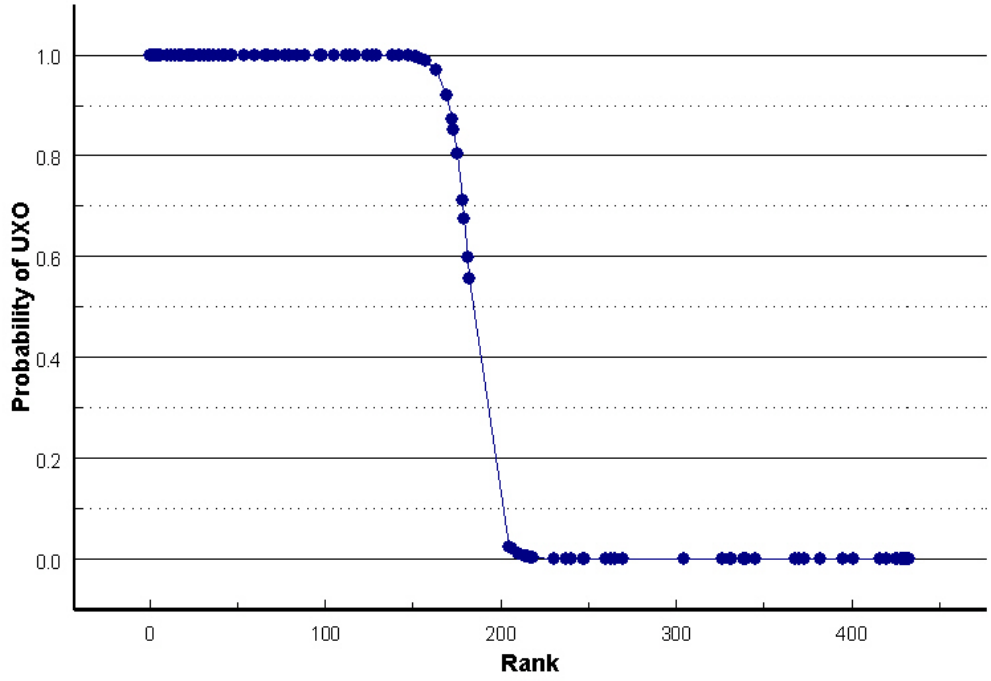
α = 31.0393
β = −0.1693

Once the parameters were derived, they may be converted into a probability of UXO as a function of rank using the following function:

Equation 5: $$P(UXO)_i = \frac{1}{1 + e^{\alpha + \beta * Rank_i}}$$

The probabilities we derived from the parameterized Equation 5 is shown on the training data in Figure 36. Figure 36 shows the probability of finding UXO as a function our prioritized dig-list rankings for the above amplitude training data. The X-axis shows high-likelihood UXO in our dig-list to the left and low-likelihood UXO to the right. The ranking shown on the X-axis is the combined ranking across all training and validation data in this step. The Y axis shows the modeled probability of finding UXO at any point on that dig-list on the Y-axis.

Figure 36. Falling probability of UXO as a function of LGP rank on training targets



6.10.2 Risk Analysis Model Applied to the Blind Data

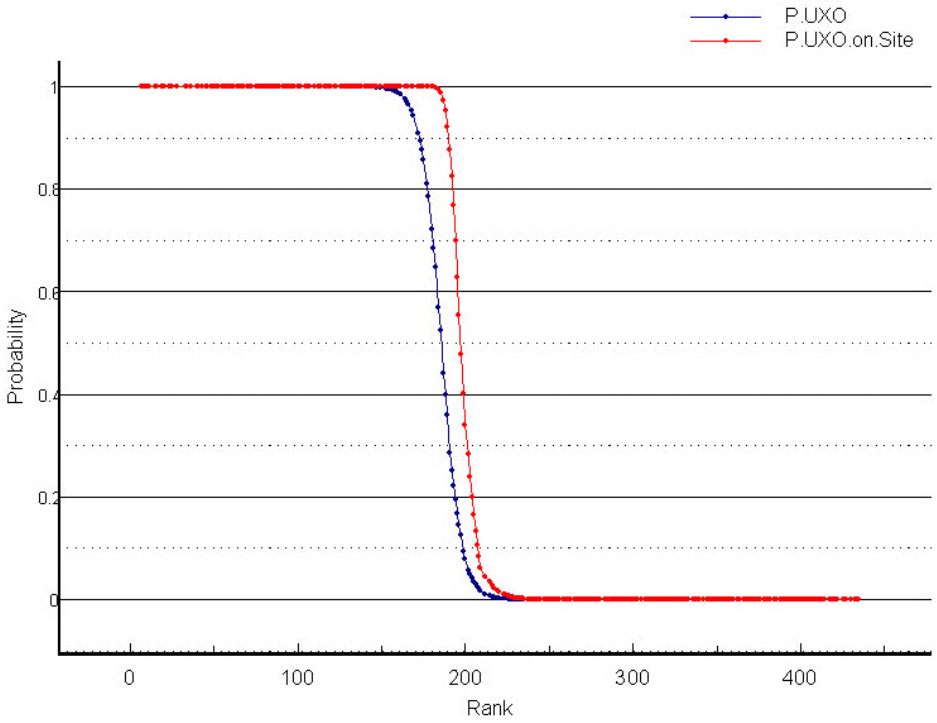
Equation 5 was then applied to the blind targets with the derived parameters, using the blind target rankings as independent variables. The resulting predicted probabilities are shown in the blue series in Figure 37. From these resulting probabilities, we then computed the cumulative probability that all blind targets ranked to the right of each ranking contained one-or-more UXO. To do so, we used "or-of-probabilities" approach described in Section 2.1.6, Equation 2 using the probabilities of all blind targets to the right of each ranking.

We then located the rank at which this cumulative probability in the tail of the UXO probability distribution fell below the designated confidence level.

The probability that more than one UXO remains in all blind-targets ranked less likely to be UXO than the current target falls below 0.025 (97.5% confidence) at ranking 217 in the combined above amplitude training and blind data. That is the 150th ranked blind target.

Figure 37 shows the computed probability that UXO remain amongst all targets ranked higher than the current target in the red series. Note that the x-axis shows the rank computed across all training and blind targets included in this step as described in Section 2.1.6.

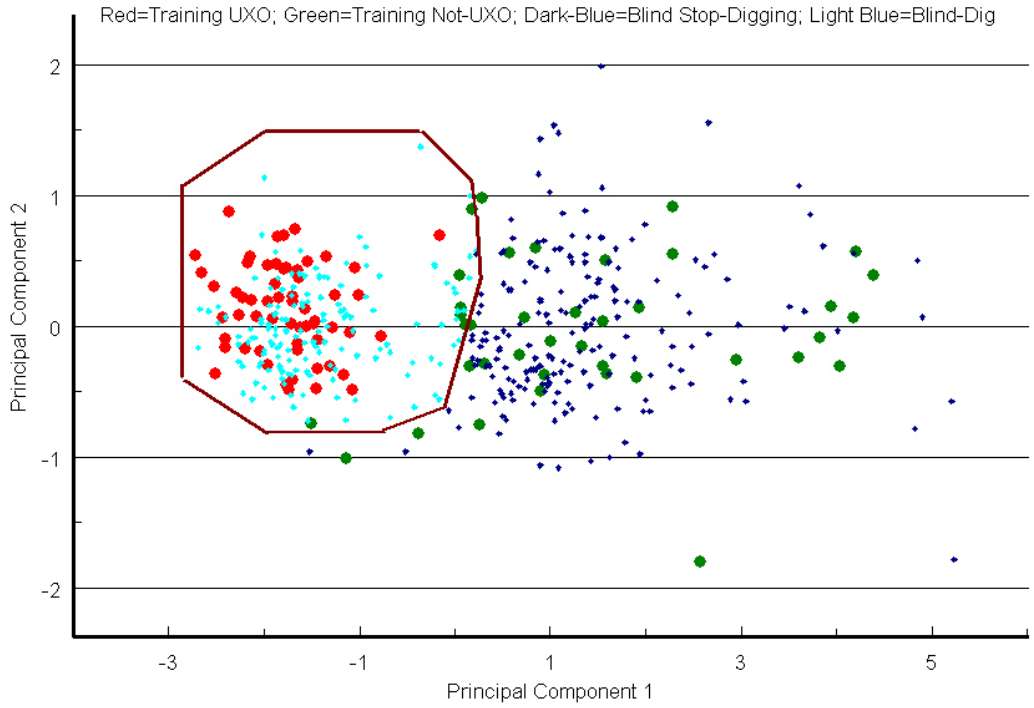
Figure 37. Probability of UXO and probability of UXO remaining on site. Blind Data



All items below rank 216 (150th blind target) have a probability that one or more UXO remains on the site of less than 0.025. Accordingly, those items will be assigned to below the stop-digging threshold.

Figure 38 is an interesting representation of these risk analysis results. To show this figure, we first converted the six attributes used in our LGP models into principal components. That figure shows the training UXO and not-UXO as red and green circles, respectively. They are shown in the attribute space defined by the two most descriptive principal components. The small light-blue dots are blind data that are *above* the stop-digging threshold set by the foregoing risk-analysis. The dark blue dots are blind data that are *below* that same stop-digging threshold. The hand-drawn polygon shows the approximate boundary between dig and not-dig.

Figure 38. Risk analysis stop-digging boundary in attribute space



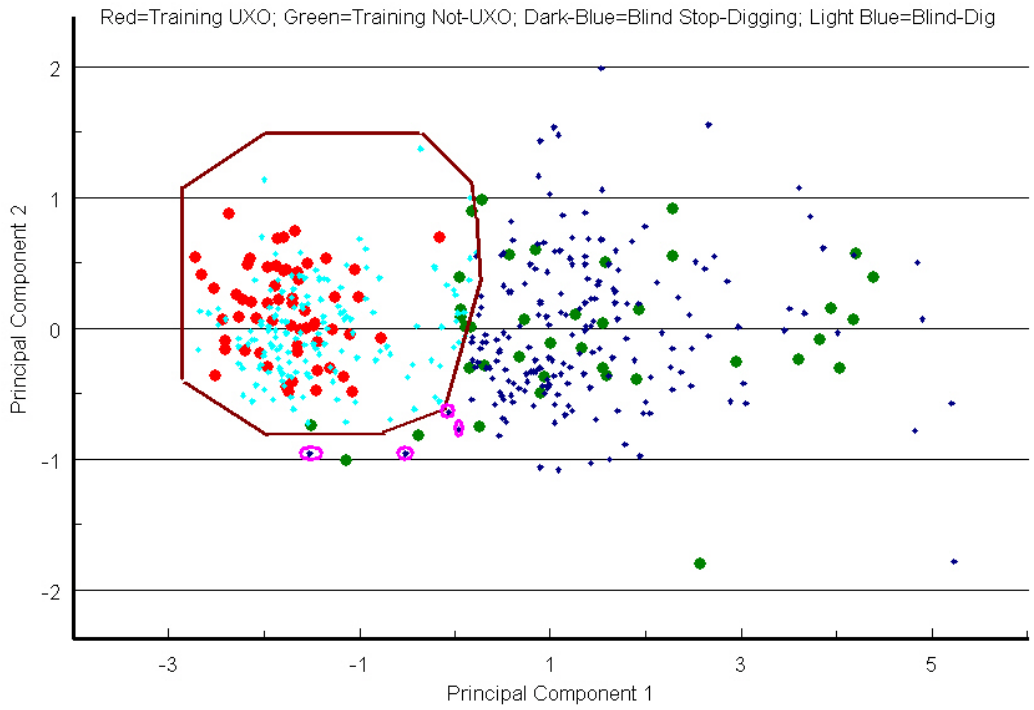
What has obviously occurred is that the risk analysis process has drawn a buffer of items that must be dug (the light blue circles) around the UXO (red circles).

6.10.3 Cannot-Analyze Targets Deriving from Risk Analysis

Note that there are some dark blue (blind targets below the stop-digging threshold) near the bottom of the polygon in Figure 38. These are on the decision boundary between dig and do not dig. But in that same region, we have only three training targets as evidence they should be left in the ground. We assess that training data density to be too low in that region and assigned the four items to cannot-analyze.

This is shown in Figure 39. That figure replicates Figure 38 but shows four targets highlighted with small magenta circles. These four targets were removed as cannot-analyze.

Figure 39. Four cannot-analyze targets caused by insufficient data density in attribute space



6.11 PRIORITIZED DIG-LIST PREPARATION

To assemble our prioritized dig-list we had to assemble the targets assessed as high probability Not-UXO by the amplitude discriminator and with all of the targets scored by the LGP ensemble predictor. We assembled below dig threshold targets together and ranked them by the probability generated for the target by the risk analysis model that assigned them to "do-not-dig." The above stop-digging threshold targets were ordered by the probability assigned to the targets by the risk analysis model that used the LGP ensemble predictor scores for ranking.

When complete, the dig-list provided a ranking, target ID, and a label whether it was above or below the stop-digging threshold or, alternatively, a cannot-analyze target as shown in Figure 40.

Figure 40. Prioritized dig-list example.

Rank	Target_ID	Comments
1	656	Above Digging Threshold
2	472	Above Digging Threshold
3	202	Above Digging Threshold
4	38	Above Digging Threshold
5	738	Above Digging Threshold
6	691	Above Digging Threshold
7	151	Above Digging Threshold
8	286	Above Digging Threshold
9	506	Above Digging Threshold
10	415	Above Digging Threshold
11	135	Above Digging Threshold

7 DATA ANALYSIS AND PRODUCTS FOR COMBINED-TRACK

This combined EM61 and MAG MTADS Track ("Combined" Track) used the statistical attributes described in Section 6.5 for the EM-only-track. There, the attributes were extracted from just the EM61MTADS DGM. In this track, they were, in addition, extracted from the MAGMTADS DGM.

The targets included in this Combined-track were all targets selected by the Program Office as an MAGMTADS target, an EM61MTADS target, or both. In other words, this track operated on the set of targets defined by:

$$Selected\_Tgt_{EM\,61MTADS} \cup Selected\_Tgt_{MAGMTADS}$$

$\cup$ is the set union operator.

Thus, the key differences between this track and the EM-only-track (previously reported), were:

1. There were more targets in this Combined-track; and
2. There were more attributes (both EM and MAG attributes were used).

This section will first describe the data used in this Combined-track and then summarize our process and results for each of those steps for the Combined-track.

7.1 DESCRIPTION OF DATA

For the Combined-track, we used all targets that had been identified by the program office as targets that had been detected by either EM61MTADS and MAGMTADS ("Combined-track targets").

We received target identification for a total of 1203 Combined-track targets. The 1203 targets are comprised of:

- 220 training (or "labeled") targets (targets for which we knew ground truth); and
- 983 blind data targets (targets for which we did not know ground truth).

Viewed another way, the Combined-track targets are comprised of:

- 713 targets that were selected by the program office as BOTH EM61MTADS targets and as MAGMTADS targets;
- 195 targets that were selected by the program office as EM61MTADS targets but not as MAGMTADS targets; and
- 295 targets that were selected by the Program Office as MAGMTADS targets but not as EM61MTADS targets.

The breakdown of the training ground truth is shown in Table 13.

Table 13. Groundtruth summary for Combined-track

Target Type	Number
UXO	59
Soils	42
Frag	32
Scrap_Metal	25
Rock	21
Halfshell	12
Nose_Frag	11
Baseplate	8
No_Contact	3
Corner_Stake	3
Survey_Point	1
Wire	1
Horseshoe	1
Wrench	1
Total	220

7.2 ATTRIBUTE EXTRACTION

The attribute extraction process from the EM61MTADS sensor has been described above. We used the same starting EM61MTADS attribute set on this track as we did on the EM-only-track ("EM Attributes").

In addition to the EM Attributes, we also extracted attributes from the MAGMTADS data. To do so, we extracted the analytic signal using Geosoft Oasis Montaj, constructed manual ellipses for them in precisely the same manner as we did for the EMMTADS targets, and extracted the same attributes from the analytic signal ellipses as we previously extracted from the EM61MTADS data. Of course, a magnetometer does not generate multiple channels of data. So there were no Ratio Statistics calculated (which presume more than one channel)

In addition, we extracted some magnetometer specific features such as the distance in meters between the high point in the positive lobe of the magnetometer signal and the low point in the negative lobe.

7.3 EXCLUDE PRELIMINARY CANNOT-ANALYZE TARGETS

We excluded some of the same cannot-analyze targets as in the EM-only-track. We did that for targets that did not have sufficiently good EM data or ellipses to discriminate on the EM61MTADS Track. That is no different on this track as they use the same EM61MTADS data as part of the data set. Those categories are described in Sections 6.4.1-6.4.4.

Another category of potential cannot-analyze targets became relevant on this track. We initially believed that the targets we were to address on this track were targets identified as

EM61MTADS *and* MAGMTADS targets. Later, it became apparent that that the requirement was to address targets that were *either*. This resulted in a large group of targets for which we had no EM ellipses or features extracted. See Section 7.1. The problem was the new (to us) MAG targets that had no EM data associated with them (295) targets. In addition, the 195 EM targets for which there was no MAG target detected posed a feature extraction problem in MAG—for the most part there was no meaningful MAG signature there. Together, these two sets of targets would have been an unacceptably high number of cannot-analyze targets.

These 295 "no-EM-features" targets had all been through a complete feature extraction process which would have been very time-consuming to emulate. The 195 "no-meaningful-MAG-signature" targets presented a related, but different problem—the likelihood that we would be analyzing data in the noise that would produce spurious signals, similar to the rut noise problem were we to extract MAG features notwithstanding the lack of a signal.

To address these issues, we noted on visual examination that the 295 "no-EM-features" targets had, overwhelmingly, very small or non-existent EM signatures. Similarly, the 195 "no-meaningful-MAG-signature" targets tended to have small EM signatures and no above-noise MAG signature. Thus, these targets appear to be very similar to the rut-noise problem targets we faced in the EM-only-track.

We addressed these two sets of targets in the same way—with an amplitude discriminator for this track. We were able to eliminate most of the 'no-EM-features" targets and most of the "no-meaningful-MAG-signature" targets as high-probability Not-UXO with the amplitude discriminator, as described below.

7.4 DERIVE AND APPLY AMPLITUDE DISCRIMINATOR

We have previously described the effect of (1) rut-noise on the EM61MTADS attributes in our discussion of the EM-only-track; and (2) the "no-EM-features" targets. We accounted for the rut-noise effect in the EM-only-track with an amplitude discriminator quite similar to the EM-only-track discriminator.

This section descries the amplitude-based pre-discriminator on the Combined-track.

7.4.1 Selecting the Amplitude-Only Attributes for the Amplitude Pre-Discriminator

We selected only those attributes from the EM attribute set that directly measure signal value. So, for example, all attributes channel to channel ratios were excluded and all high-level attributes measuring signal-decay were excluded. In addition, we selected only "circle" ring features. These features do not require an ellipse and therefore greatly compressed the feature extraction for the attribute discriminator.

For this track, we started by measuring the mutual information between the training target labels and the various binned amplitude attributes.

The two attributes with the highest level of mutual information with the training target labels were as follows:

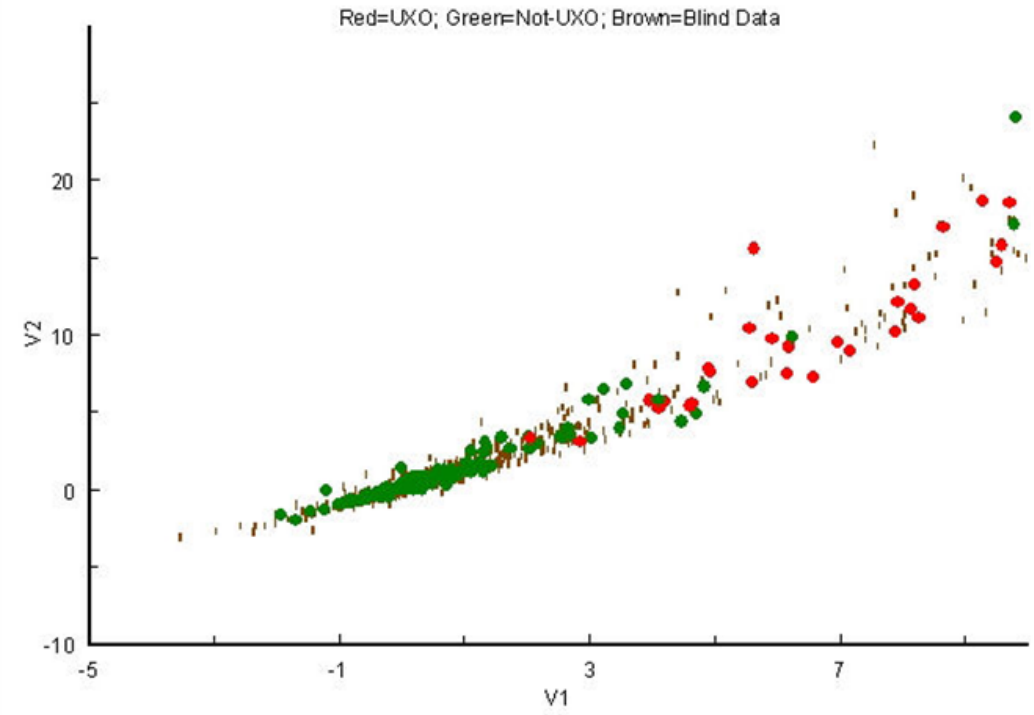
- COMAMP-V.1: The Channel 3 (final decay channel) Signal Value in the outer region of the target ellipse. The mutual information of this attribute with the training labels was 0.59.

- COMAMP-V.2: The Channel 3 (final decay channel) Signal Value in the next-most outer region of the target ellipse. The mutual information of this attribute with the training labels was 0.58.

Those two were selected as the basis for the amplitude-based discriminator. These turned out to be very similar to the features selected by mutual information on the EM-only-track.

We began our analysis of these features by visually inspecting the selected attributes and how well they segregate Not-UXO. Figure 41 shows the selected features and how well they discriminate the low-amplitude Not-UXO from UXO. COMAMP-V1 is shown on the X-axis and COMAMP-V2 is shown on the Y-axis.

Figure 41. Closeup of amplitude features for Combined-track on training and blind data. X-axis is COMAMP-V1 and Y-axis is COMAMP-V2.



On these two attributes alone, we have good class separation for low-signal-value targets. The lowest ranked UXO is at approximately 2.7 on COMAMP-V1. And, the distribution of the blind data (the small brown dots) matches the training data quite nicely.

We converted these two attributes into a single, best feature using principal components analysis. Effectively, the first principal component on these data projects each target onto the best regression line fitting the data, which is exactly what we want.

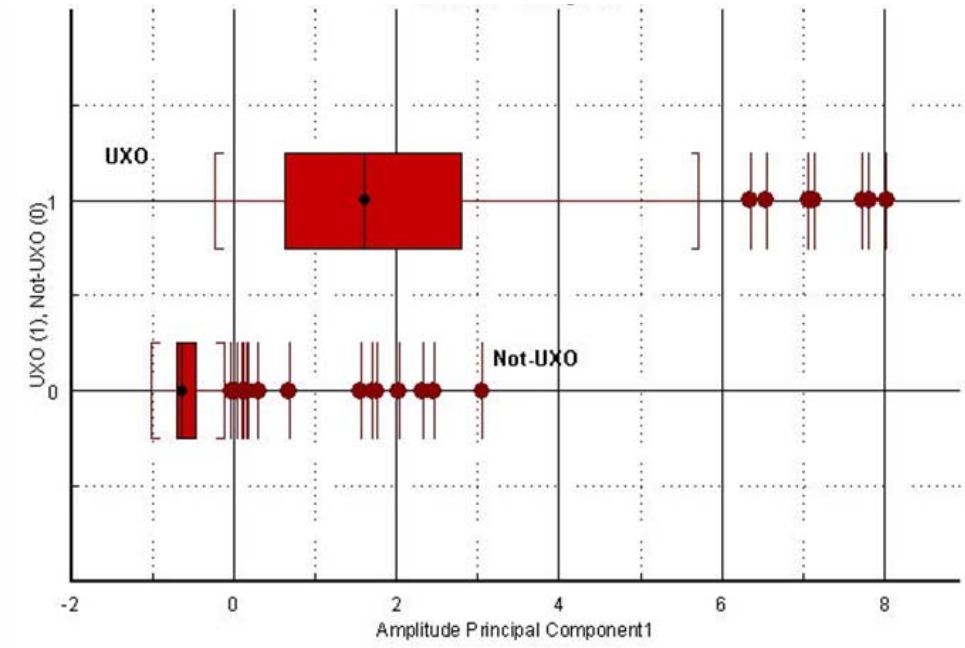
Accordingly, we performed principal component analysis on COMAMP-V1 and COMAMP-V2 and used the first principal component. The principal component used ("Amplitude Principal Component 1") may be described as follows:

- COMAMP-V1 is normalized with a mean of 5.03 and a standard deviation of 10.47.
- COMAMP-V2 is normalized with a mean of 3.09 and a standard deviation 6.40.

- Amplitude Principal Component 1 is 0.71 * Normalized COMPAMP-V1 + 0.71 * Normalized COMPAMP-V2.

At this point we have reduced the amplitude attributes to a single attribute (Amplitude Principal Component 1), which, by itself provides a ranking. That is, the higher the value of Amplitude Principal Component 1, the more likely an item is to be UXO. This is demonstrated in Figure 42, which shows that this component, by itself, very effectively segregates a portion of the non-UXO from UXO.

Figure 42. Comparative distribution of UXO and Not-UXO on Amplitude Principal Component 1. Training targets only.



The shaded boxes in Figure 42 show the inter-quartile range for each target type (UXO or not-UXO). The brackets show the range of values that are not outliers and the circles show outliers.

It is apparent that the great bulk of the UXO is concentrated between Amplitude Principal Component 1 of 0.7 and 4.8. The lowest ranked UXO by this metric is Amplitude Principal Component 1 value equals -0.23. On the other hand, about 75% of the non-UXO have a component value below -0.23.

Accordingly, this component provides a good basis for performing elimination of targets as high-probability MEC based on amplitude measurement alone.

Furthermore, this component provides a highly statistically significant split of the training data into UXO and Not-UXO. The lowest component value for any UXO is -0.23. If we split the training data at Amplitude Principal Component 1 < -0.23, we obtain the following 2x2 contingency table for UXO and Not-UXO above and below the split:

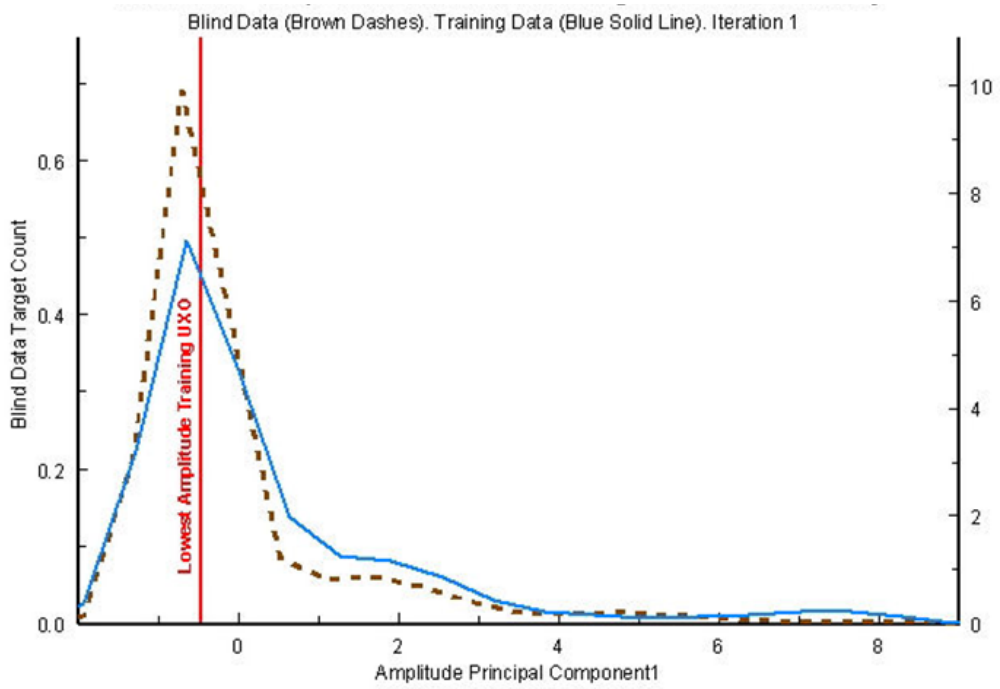
Table 14. Two-by-two contingency table for Combined-track Amplitude Principal Component 1 as a Discriminator

	Below Split	Above Split
UXO	0	59
Not-UXO	128	23

The Chi Square statistic for this table is computed with one degree of freedom using Yates Continuity Correction. The probability of Chi Square for this table is 0.000. Accordingly, we conclude that the split of the training data at -0.23 using Amplitude Principal Component 1 produces a highly statistically significant separation of Not-UXO from other targets.

We then checked that the distribution of the training and blind data on the selected component provided a reasonable match. Thus, we analyzed the density of the training and blind data as a function of Amplitude Principal Component 1. That analysis is shown in Figure 43.

Figure 43. Density of Amplitude Principal Component 1 on training and blind data for Combined-track.



The match between the densities of the training and blind data is quite close. This suggests that the training data is reasonably representative of the blind data on this attribute. Accordingly, we are comfortable generalizing our discrimination on the training data to the blind data using this component.

7.4.2 Assigning Targets to High-Confidence Not-UXO Based on Amplitude Principal Component 1

The next task in this process was to determine where, on the Amplitude Principal Component 1 axis, we may safely say that the probability that all items with lower Amplitude Principal Component 1 are Not-UXO. To do that, we turned to residual risk analysis methodology.

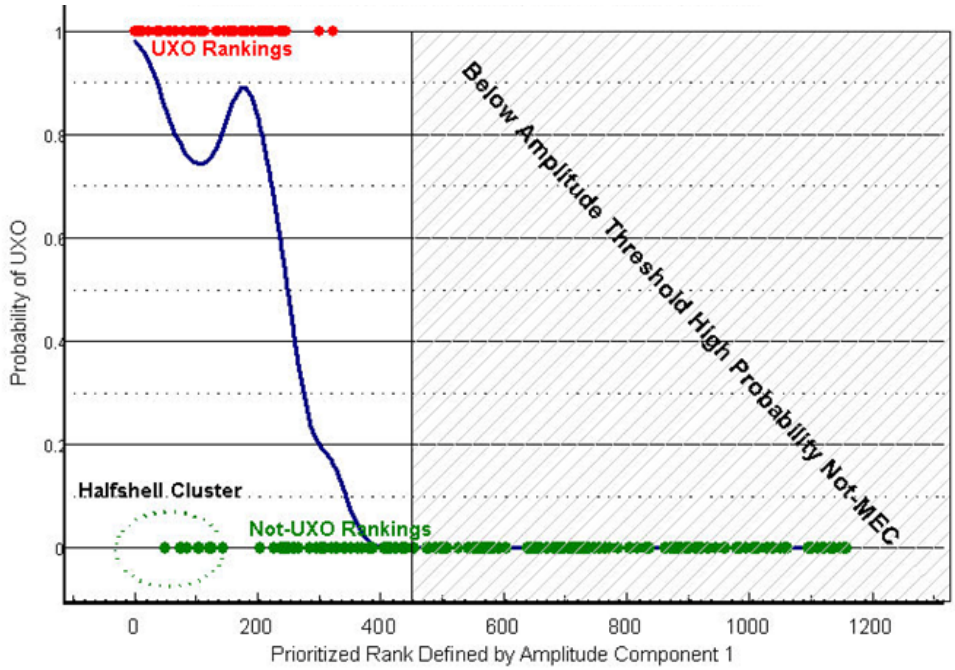
We first converted Amplitude Principal Component 1 values into ranks across the entire training and blind data sets. In making this conversion, lower values of Amplitude Principal Component 1 were interpreted as higher rank (that is, less likely to be UXO). We then evaluated Logistic regression, exponential regression, power law regression and kernel regression.

The first three functional types were deemed inappropriate because of the local ups and downs of the probability as a function of Amplitude Principal Component 1 (see Figure 44). Kernel regression, on the other hand, does a good job of modeling such local irregularities and is generally preferable to the others, all other things being equal, because it is a single-parameter model. Accordingly, we used kernel regression with a Gaussian kernel as set forth in Equation 3.

We derived an optimal value for the width parameter, α, in Equation 3 using leave-one-out cross-validation on the training data in the manner as described in Section 6.8.4. The value determined for the parameter, α, is 26.593.

Next, we applied the above Gaussian kernel, generated by the training data, using the derived kernel width α parameter, to the ranked blind data. This generated a probability that each blind data item is UXO. Figure 44 shows that probability as a function of rank.

Figure 44. Probability of UXO as a function of Amplitude Principal Component 1 rank on training data.



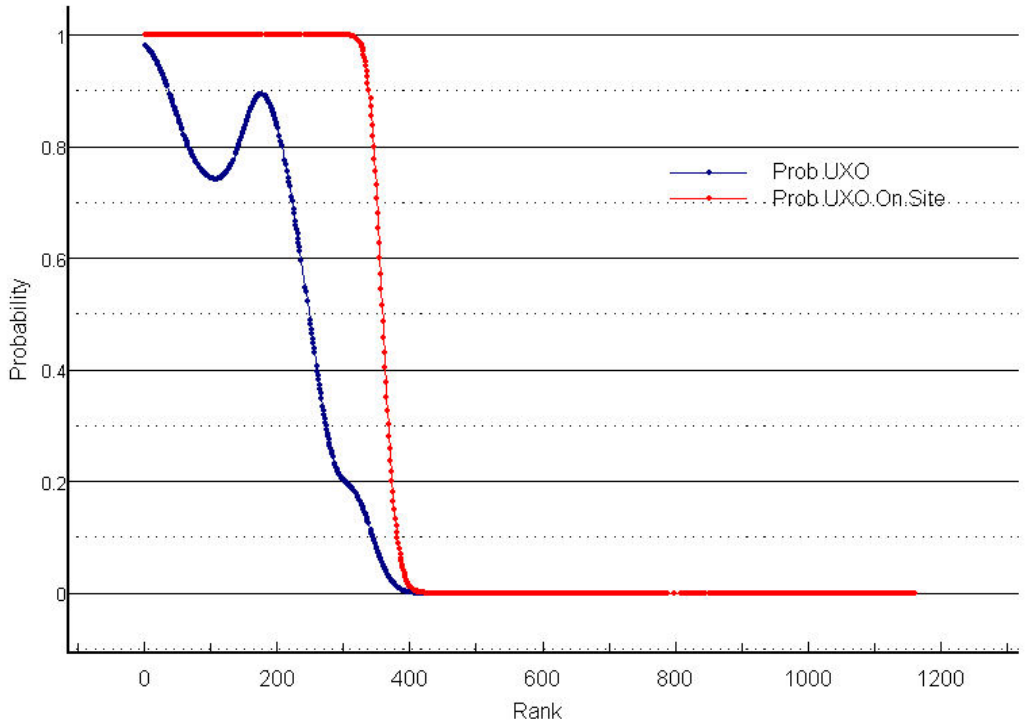
The red circles shown in Figure 44 are the UXO as ordered by the Amplitude Principal Component 1 rankings (the rankings shown on the x-axis are rankings across the entire training

and blind data sets, combined). The green circles are the not-UXO, ordered by Amplitude Principal Component 1 rankings. The blue line is the local probability generated by kernel regression that a given rank is UXO. Note that at around ranking 180, the probably of UXO increases. The reason for this is the circled cluster of half-shells that the amplitude component rankings find very early. The gap between that cluster and the remaining Not-UXO found causes the bump in the local probability value.

Once those probabilities were predicted on the blind targets, we then assessed the probability that all blind targets ranked above each Amplitude Principal Component 1 ranking contain one-or-more UXO (note again that higher rankings correspond to lower values of Amplitude Principal Component 1) using the "or-of-probabilities" approach described in Section 2.1.6, Equation 2, applied to all such higher ranked targets. This generates the residual risk cumulative probability that one-or-more UXO remain on site at each ranking.

Figure 45 shows the result of that computation—both the probability of UXO and the probability of UXO remaining on site for the amplitude discriminator for the Combined-track.

Figure 45. Kernel regression of probability of UXO and probability of UXO remaining on site as a function of Amplitude Principal Component 1 rank. Blind data projections.



The blue line in Figure 45 is the modeled probability of UXO as a function of Amplitude Principal Component 1 rank. The red line is the cumulative probability that one or more UXO remains in any target ranked to the right of the plotted rank. When the red line falls below the critical p value, we assess all targets remaining to the right of that value as high-probability Not-UXO.

The critical value we used was the Bonferonni corrected p-value for a 95% confidence level. We must use the corrected value because we are using two different discriminators on this track and each does a probabilistic assessment.[27] Properly corrected, the critical value here is $p \leq 0.025$.

The probability of any UXO remaining to the right of the measured ranking falls below 0.025 at ranking 446. This is equivalent to a determination that any target with an Amplitude Principal Component 1 value of less than or equal to -0.4857 is high-probability Not-UXO in the Combined-track.

Using this criterion, one-hundred eleven training targets fell into the high-probability not-UXO region. Six-hundred six blind targets fell into the high-probability not-UXO region. These targets were excluded from the attribute reduction, LGP modeling and subsequent residual risk analyses as high-probability not-UXO.

Once done, we revisited the 295 "no-EM-features" targets mentioned in Section 7.3. (That is, the 295 program office MAGMTADS targets that were NOT also detected as EM61MTADS targets.) After the amplitude discriminator had been applied, only twelve training and thirty-five blind targets remained as possible UXO out of the original 295. Those forty-seven targets were excluded as cannot-analyze.

In addition, at this point, we revisited the 195 "no-meaningful-MAG signature" targets mentioned in Section 7.3. (That is, the 195 program office EM61MTADS targets that were not also detected as MAGMTADS targets.) After the amplitude discriminator had been applied, only six blind targets remained as possible UXO out of the original 195. Those six targets were excluded as cannot-analyze.

7.5 ATTRIBUTE REDUCTION ON ABOVE AMPLITUDE TARGETS

From this point on in the Combined-track, we operated on only targets that fell above the amplitude discriminator threshold.

The entire EM and MAG attribute sets described elsewhere were our starting point on attributes. For these data, we had 87 training data instances remaining after the amplitude pre-discriminator and after removal of cannot-analyze targets. Our starting rule of thumb for attribute selection is that we should have no more than one attribute for every ten rows of training data. Thus, in this case, we want to select, at most, eight or nine attributes for training. This section describes how we reduced the large number of attributes we started with to just a few highly relevant attributes for this track.

There are a number of different approaches to determine which attributes and which set of attributes have the most predictive power with respect to a target output. Two of the more commonly used methods are correlation and mutual information.[28] Both were used in this step and both produced a small subset of useful features for further analysis.

---

[27] See: http://mathworld.wolfram.com/BonferroniCorrection.html.

[28] See, e.g.: Hall, Mark, "Correlation-based Feature Selection for Machine Learning." Doctoral Dissertion. University of Waikato, Hamilton NZ, 1999 (CFS Subset Evaluation evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them; subsets of features that are highly correlated with the class while having low inter-correlation are preferred.); and Hanchuan Peng, Fuhui Long, and Chris Ding, "Feature selection based on mutual information: criteria of max-

7.5.1 First Order Attribute Analysis

Our first step in attribute reduction is to determine which attributes have the highest level of mutual information with the labels of the training data as UXO or Not-UXO. To measure mutual information, numeric data must be discretized into bins. Accordingly, we split each attribute into ten equal frequency bins. Once the attributes were binned, we then measured the mutual information between each attribute and the labels of UXO vs. Not-UXO.

One issue became immediately apparent. Across the board, EM attributes have a higher level of mutual information with the target labels than do the MAG attributes. The difference is pronounced. For example, we ranked all of the attributes in order of their mutual information with the class labels. The top ranked MAG attribute ranked number 90 in the overall data set.

Table 15 demonstrates this. It was performed with 42-fold cross-validation and the numbers shown are the average across all folds. The three Mag attributes with the highest level of mutual information with the target output are all ranked lower than the best 89 EM attributes.

Table 15. Relative ranking of best EM and Mag attributes

Attribute Description	***Average Mutual Information Rank amongst all Attributes***	***Average Mutual Information with Target***
Best EM61 Attribute	1.2	0.601
2nd Best EM61 Attribute	1.9	0.583
3rd Best EM61 Attribute	3.5	0.551
Best MAG Attribute	90.5	0.281
2nd Best MAG Attribute	105.8	0.267
3rd Best MAG Attribute	150.8	0.235

7.5.2 Subset-Based Attribute Selection

Had we stopped with just our first-order analysis, we would have excluded all MAG attributes and the Combined-track would have looked much like the EM-only-track, just on a different set of targets. This section describes how we evaluated subsets of attributes, instead of just one attribute at a time, to select a parsimonious and highly predictive attribute set.

Although high mutual information with the target labels is a desirable quality in an attribute, building an attribute *set* that is both effective and parsimonious is a somewhat more complex process. The reason is, a desirable attribute set contains (1) Attributes with a high level of mutual information or correlation with the target labels (high relevance); and (2) The information

---

dependency, max-relevance, and min-redundancy," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, pp.1226-1238, 2005 (MRMR selects attribute subsets based on a high level of mutual information with the target output and a low level of mutual information amongst the attributes in the selected attribute subset).

contained in each selected attribute should provide DIFFERENT information about the target labels than do the other attributes (low redundancy amongst the selected attributes).[29]

Thus, two attributes that carry exactly the same information about the target labels are no better than either one of the attributes singly. By themselves, correlation and mutual information only measure the relationship between a single attribute and the target output. In other words, correlation and mutual information, by themselves, frequently do not produce a good attribute subset by themselves.

To evaluate an entire attribute subset, with respect to the output, we used the algorithms referred to in Footnote 28. They are "CFS Subset Evaluation," which uses correlation as the measure of relevance and redundancy for an entire data set and "MRMR" which used mutual information as the measure of relevance and redundancy. We use both measures in this track.

Despite the much lower mutual information level of the Mag attributes, our conjecture was that the MAG attributes, having been collected by a different sensor technology, would contain different information about the class labels than the EM attributes. Accordingly, our initial cut on attributes included:

(1) The best attribute set selected by a semi-greedy, Best-First algorithm with backtracking attribute selection algorithm, using CFS Subset Evaluation. We used 50-fold cross-validation and selected all attributes that were selected in more than 50% of the folds. We selected CFS for this step in the hope that it would uncover important MAG attributes that mutual information ranking by itself did not uncover (see Table 15). However, the attribute set selected by this algorithm was comprised of nine EM attributes. No MAG attributes were selected; plus

(2) The eleven top ranked MAG features using the MRMR algorithm.

We refer to this attribute set, containing twenty attributes, as "Combined Attribute Set 1."

7.5.3 Feature Exclusion using Tree Ensemble

Combined Attribute Set 1 contains twenty features. This is more features than desirable given the training set size as noted above. Our next step in feature reduction was to use an ensemble of decision trees to reject a portion of the attributes.

Accordingly, we used Combined Attribute Set 1 as inputs to an ensemble-based decision tree algorithm called Random Forests ("RF"). RF is not particularly good at identifying parsimonious attribute subsets. It is, however, quite good at quickly excluding attributes as being of no, or marginal, relevance. For this step, we use the Gini Variable Importance measure generated by the RF algorithm. We reviewed the variable importance metric generated by the package and there was a significant break in importance after the twelfth ranked attribute from COMBINED Attribute Set 1.

Accordingly, we excluded the bottom-ranked eight attributes from Attribute Set 1 and created Combined Attribute Set 2 as shown in Table 16. The factors influencing this cutoff were: (1)

[29] Hanchuan Peng, Fuhui Long, and Chris Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, pp.1226-1238, 2005.

This was the first substantial break that occurred after four MAG attributes had been selected; (2) The eight excluded variables had a Gini measure less than one-tenth of the most important attribute; and (3) After exclusion of eight variables, we are much closer to our goal of having no more than eight attributes for training and we have not excluded at least some of the MAG attributes.

Table 16. Reduction of Attribute Set 1 using Random Forests to Exclude Attributes

Attribute Type and Rank	*Gini Variable Importance*	*Excluded from Further Analysis?*
EM Attribute 1	16.351	No
EM Attribute 2	16.209	No
EM Attribute 3	12.552	No
EM Attribute 4	10.305	No
EM Attribute 5	8.264	No
EM Attribute 6	8.231	No
EM Attribute 7	6.69	No
EM Attribute 8	5.389	No
Mag Attribute 1	4.129	No
Mag Attribute 2	2.726	No
Mag Attribute 3	2.092	No
Mag Attribute 4	1.628	No
Mag Attribute 5	1.158	Yes
Mag Attribute 6	0.981	Yes
Mag Attribute 7	0.978	Yes
Mag Attribute 8	0.973	Yes
Mag Attribute 9	0.565	Yes
Mag Attribute 10	0.408	Yes
Mag Attribute 11	0.284	Yes
EM Attribute 9	0.087	Yes

7.5.4 Final Attribute Reduction Using LGP and Visual Inspection

Attribute Set 2 included eight EM attributes and four MAG attributes. The four MAG attributes were the last four ranked attributes in this approach.

We then used Attribute set 2 as inputs to a ten-fold LGP cross-validation run and examined the frequency with which each of the items in Attribute Set 2 appeared in the thirty best programs across all LGP runs in the cross-validation. The results are show in Table 17.

Table 17. Attribute reduction using LGP attribute frequencies

Attribute Description	*Frequency*
EM Attribute 6	0.975
EM Attribute 3	0.863857143
MAG Attribute 2	0.850428571
EM Attribute 2	0.588428571
EM Attribute 1	0.546428571
EM Attribute 8	0.381142857
EM Attribute 7	0.338285714
EM Attribute 5	0.327857143
EM Attribute 4	0.201428571
Mag Attribute 3	0.171
Mag Attribute 1	0.073142857
Mag Attribute 4	0.063714286

Using the same naming convention as in previous tables for this track, Table 17 shows the attribute description and the frequency with which that attribute was selected as important by the best LGP programs. A value of 0.975 in the Frequency column means that 97.5% of the best programs generated by the LGP cross-validation runs included this attribute.

Before the final attribute selection, we removed outliers from the attribute space of the top five attributes identified by the LGP frequency analysis (see Table 17). The purpose of this, of course, is to tailor the cannot-analyze targets to the specific attributes we will be examining and to identify areas in attribute space where the training data does not sufficiently define the blind data. This process was performed by visual inspection of attribute space.

Figure 46. Example of assignment of targets to cannot-analyze based on attribute space outlier analysis

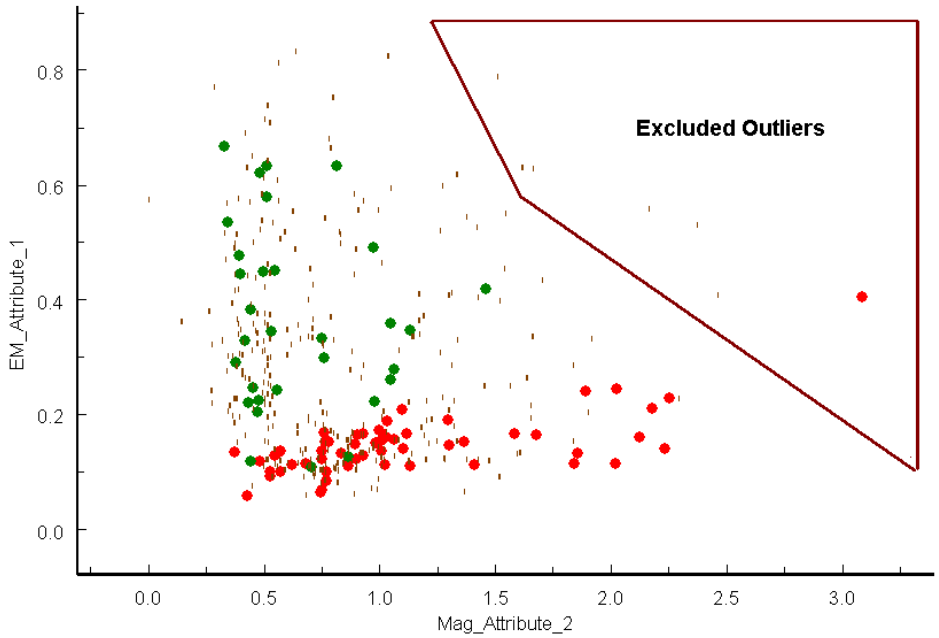


Figure 46 demonstrates this process of defining outliers as cannot-analyze targets. It is a plot of two of the five attributes identified as possible attributes for final modeling. The polygon shows the training and blind targets we assessed as outliers. All targets in that polygon were excluded as cannot-analyze. Based on this plot, the following targets were assigned as cannot-analyze targets (from left to right): Targets 780, 1328, 976, 810, 840, 878, and 703.

Having removed the outliers, we then serially examined graphics of subsets of the attribute space starting with the top two attributes. The graphic was a plot of the two most important principal components of the attribute space. At each step, we added the next most frequent attribute defined in Table 17 and redrew the graph. At each step, we examined the attribute space for the quality of the separation of the UXO class and the Not-UXO class. We continued to do so until no further improvement in the separation of UXO from Not-UXO occurred.

Figure 47, Figure 48, and Figure 49 show the attribute space charts used for each step in this analysis.

Figure 47. Two most frequent LGP identified attributes—a principal components view of attribute space. EM Attribute 6 and EM Attribute 3.

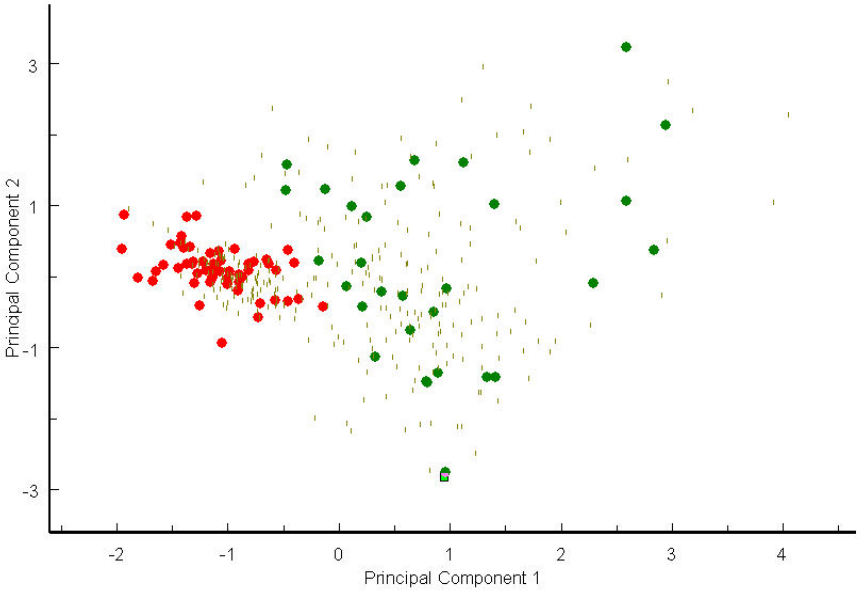


Figure 48. Three most frequent LGP identified attributes--principal components view of attribute space. EM Attribute 6, EM Attribute 3, and Mag Attribute 2.

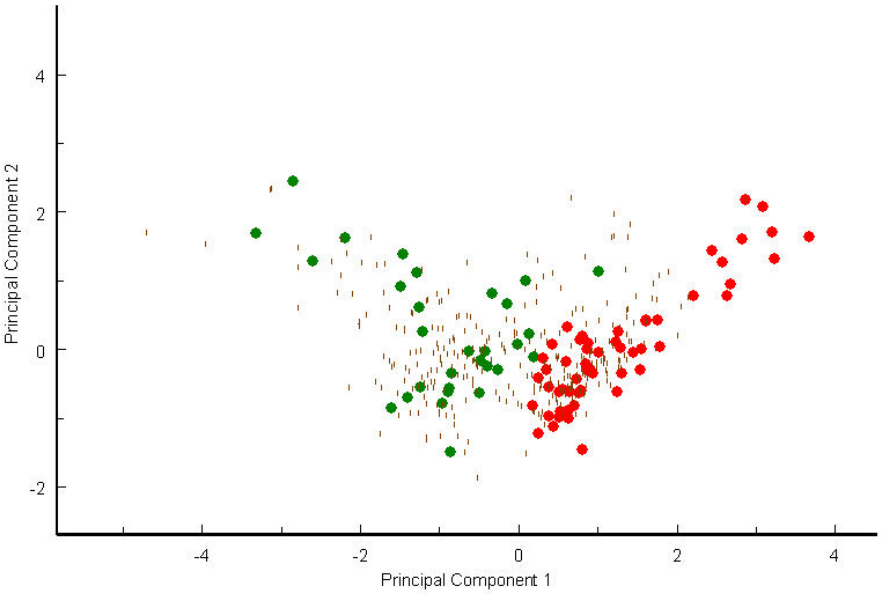
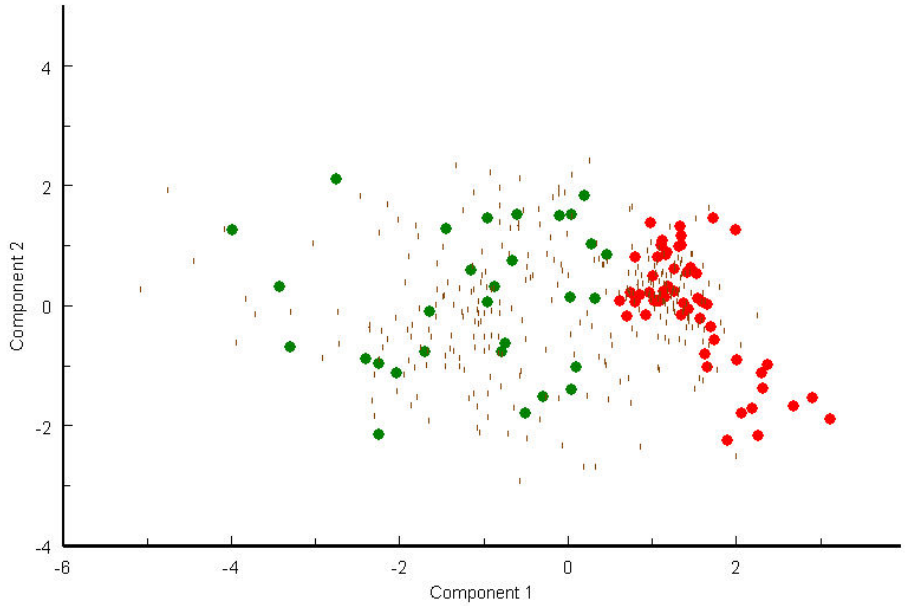


Figure 49. Four most frequent LGP identified attributes--principal components view of attribute space. EM Attribute 6, EM Attribute 3, Mag Attribute 2 and EM Attribute 2



The foregoing three figures show the effect of adding one variable at a time to attribute space in the variable frequency order evaluated by the LGP algorithm. Two notes on the foregoing:

1. Comparing Figure 47 and Figure 48, it is not clear that adding Mag Attribute 2 provides any improvement over using just the first two attributes (EM Attribute 6 and EM Attribute 3). We elected to retain it nevertheless as it is the only MAG attribute remaining at this point in the attribute selection process. As noted in the next paragraph, we achieve complete linear separation by adding only one additional attribute beyond Mag Attribute 2. So the final attribute set will have only four attributes.
2. There is no point in going further than the four attribute set shown in Figure 49. The first principal component of the four attribute set achieves complete linear separation of the two classes. That is, there is a linear transform from the four dimensional attribute space shown in Figure 49 to a single vector thru that space that perfectly classifies all the training data when all training data is projected onto that vector.

Accordingly, we used two final attribute sets:

1. Attribute Set 3 on this track was comprised of :
    a. EM Attribute 6
    b. EM Attribute 3
    c. Mag Attribute 2
    d. EM Attribute 2.
2. Attribute Set 3PC on this track was comprised of the two principal components shown in Figure 49.

7.6 LGP DISCRIMINATION OF UXO vs. NOT UXO

After removing the high-confidence Not-MEC from the training and blind data in the amplitude discriminator step and the cannot-analyze targets as described above, we were left with 85 Training and 298 Blind targets. This section describes how we applied the LGP Classifier to these reduced data sets.

LGP Discrimination took place in two steps: (1) Cross-validation to set the noise parameter; and (2) Bagging to produce a model and prioritized dig-list.

7.6.1 Cross-Validation to Set the Noise Parameter

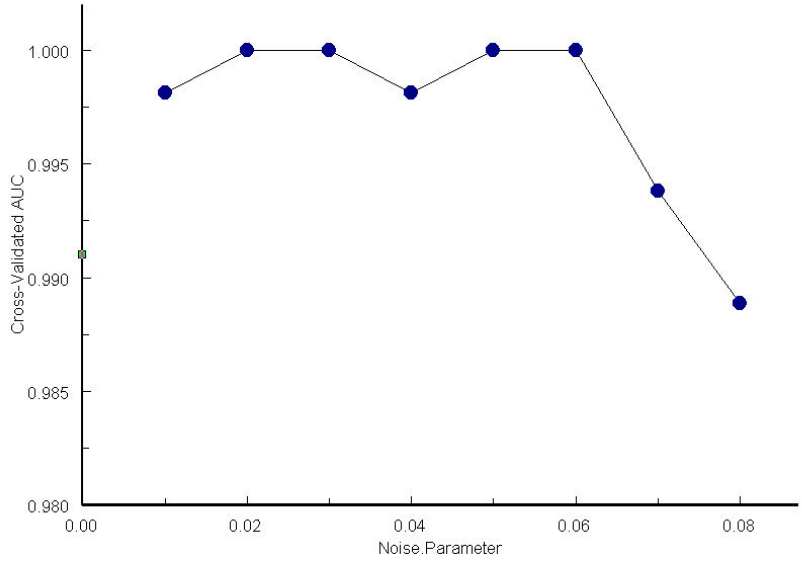
This is a small training set. To prevent over-fitting the training data, we added a small amount of Gaussian noise to the inputs. The standard deviation of the added noise is set attribute by attribute. A noise parameter of 2% means that the standard deviation of the Gaussian noise is set to 2 percentiles of the distribution of that variable.

Setting the amount of noise is an empirical process dependent on the data set at hand. We set the noise parameter using ten-fold cross validation, testing noise settings of 1% thru 8% in increments of one. In performing the cross-validation, the default settings of Discipulus™ LGP software were used with the following exceptions: (1) The fitness function used was Area under the curve; (2) The termination criterion for each run was 40 generations without improvement; (3) The number of runs performed in each project was 20 runs. Of course, the noise level was varied for parameter selection.

Most noise settings produced an Area under the curve (AUC) summed over the held-out cross-validation data of 0.99 or better (a very good ROC curve).

Figure 50 shows the cross-validated AUC over all tested noise settings on Attribute Set 3. The four best noise parameter settings were 2, 3, 5, and 6% (AUC=1.0). We selected a single noise parameter setting of 5.5% noise for further analysis as it is halfway between two of the best cross-validated values.

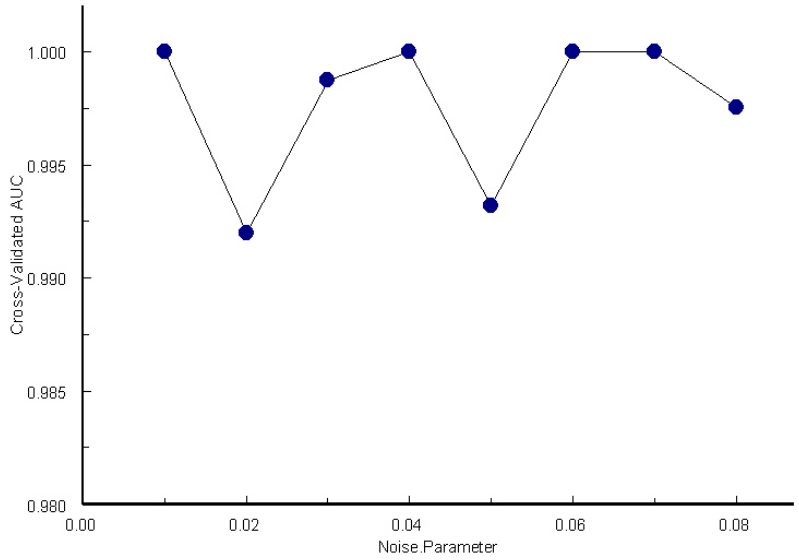
Figure 50. Cross-validated area under the curve for various noise parameter settings on attribute Set 3.



The number of misranked Not-UXO as a function of noise level followed the same pattern as Figure 50 and provided the same decision support so it is not reproduced here.

Figure 51 shows the cross-validated AUC over all tested noise settings on Attribute Set 3 PC. The four best noise parameter settings were 4%, 6%, and 7% (AUC=1.0). We selected a single noise parameter setting of 6.5% noise for further analysis on this attribute set as it is halfway between two of the best cross-validated values.

Figure 51. Cross-Validated area under the curve for various noise parameter settings on Attribute Set 3PC.



7.6.2 Bagging to Produce the LGP Ensemble Model

To prepare the prioritized dig-list, we performed 30 bagging runs at the selected noise parameter setting for each attribute set selected.

The training data for each "bag" is selected by taking n samples (each sample being a specific training target together with all attributes and labels associated with that target) with replacement from the full training data set, where n is equal to the number of training data points. The training targets NOT selected for that "bag" (about 32% of the training data in each "bag") are not used in training for that "bag". Rather, they are held-out from training process. These "held-out" training targets are referred to as the "out-of-bag" data.

The default settings of Discipulus™ LGP software were used with the following exceptions: (1) The fitness function used was area under the curve; (2) The termination criterion for each run was 40 generations without improvement; (3) The number of runs performed in each project was 20 runs. Each project used a different random "bag" for the training data.

Our final model was, therefore, an ensemble of sixty LGP Evolved Programs thirty on attribute set 3 with a noise level of 5.5% and thirty trained on attribute set 3PC with a noise level of 6.5%. We refer to this as the LGP ensemble predictor

7.6.3 Out-of-Bag Error to Estimate Performance on Blind Data

Predictions on the out-of-bag data are used to predict the expected error on the blind data and for residual risk analysis. They are used because the labels on the out-of-bag data are unknown to the LGP algorithm when it is training. Thus, the out-of-bag error is our best estimate of the expected error (1-AUC) on blind data.

We computed the out-of-bag error as follows: Each training target has multiple predictions from the ensemble model produced when that target was in the out-of-bag data. Those predictions are summed for each training target and averaged. This average was treated as our prediction for that training data point. The predictions, of course, permit us to rank the training data points relative to each other in a prioritized dig-list. That list produces a ROC Chart.

The out-of-bag ROC chart on this track is easy to summarize. All Not-UXO were ranked by LGP below all UXO. The area under the curve of the ROC chart for these results on the out-of-bag training data is, therefore, 1.0. As these error predictions are on unseen, out-of-bag data, we expect similar numbers for the blind data.

7.6.4 Scoring the Blind Data with LGP Models

We then score the blind targets using the same LGP ensemble predictor. The score for each blind target was the average of all outputs from the models in the ensemble for that target.

7.7 RESIDUAL RISK ANALYSIS FOR LGP MODELED TARGETS

This section describes the application of our risk analysis methodology to the LGP ensemble predictor described in the previous section for the Combined-track.

In summary, we took the scores of the LGP ensemble predictor for both training and blind data for this step and assembled them to produce a combined ranking across both data sets. In making that conversion from scores to ranks, a low LGP score was converted to a high ranking (that is, a low LGP score translates to a ranking that is less likely to be UXO). Then, we built a regression model of the probability of UXO as a function of the combined rank, using that combined rank and the known groundtruth for the training data. Finally, we applied that regression model to the blind data and calculated the residual risk from the resulting probabilities for the blind targets

After assembling the combined ranks for this track, the next step in risk analysis was to build a probabilistic regression model of the UXO/Not-UXO groundtruth as a function of the rank across the training and blind data in this step. To build the model, we used the training data and associated groundtruth labels.

The four functional forms we considered for risk analysis were: exponential fit, power law fit, logistic fit and kernel regression. We discarded exponential or power law fits to model probability in this track. Both are monotonically decreasing functions with a continuously increasing first derivative. The perfect ranking on the training data in this track was better represented a step-like function. Accordingly, the obvious functional form to use here was a logistic function derived using logistic regression, which inherently has a step-like shape.

Like the EM-only-track, this track also produced a perfect ranking on the training data. So we had numeric issues on this track similar to the ones described for the EM-only-track in Section 6.10.1. We solved those numeric issues in the manner described in that section.

Having solved the numeric issues, we then performed logistic regression, which optimizes two parameters in the functional form shown in Equation 4. The following values were derived for these two parameters:

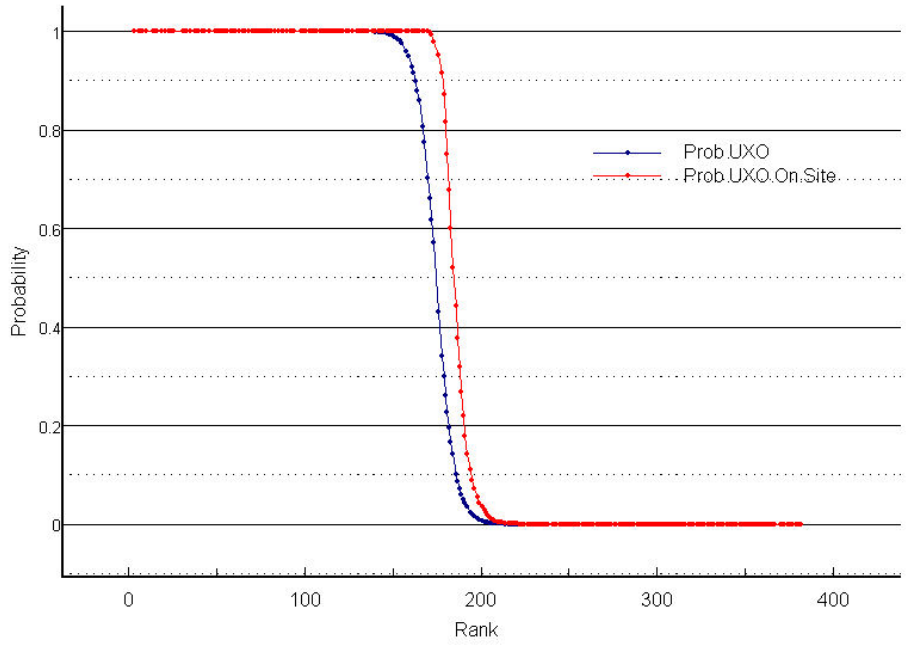
$$\alpha = 33.06633$$

$$\beta = -0.18948$$

Then, we substituted these parameter values into Equation 5 to predict probabilities of UXO on the blind data by rank using the ranks derived from the blind LGP ensemble predictor scores as the independent variable. These probabilities are shown in the blue line in Figure 52 for the blind targets remaining at this point in the Combined-track.

Once we derived these probabilities for each blind target, we calculated for each rank, the cumulative probability that one-or-more of the blind targets that have a higher ranking than the rank for which we are making the calculation contain UXO. Those cumulative probabilities are calculated using the "or-of-probabilities" approach described in Equation 2 in Section 2.1.6. These cumulative probabilities that UXO remains on the site are shown in the red line in Figure 52 for the blind targets remaining at this point in the Combined-track.

Figure 52. Residual Risk Analysis for LGP models on Combined-track. Blind data.



When the red line reaches a critical p value, we assess all targets remaining to the right of that value as high-probability Not-UXO.

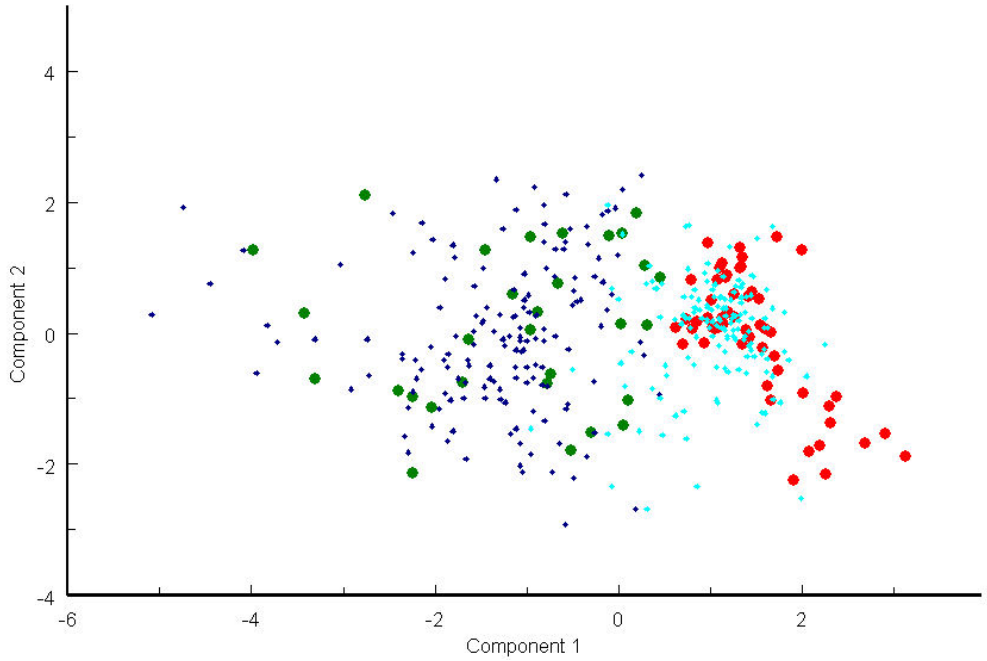
The critical value we used was the Bonferonni corrected p-value for a 95% confidence level. We must use the corrected value because we are using two discriminators on this track.[30] Properly

[30] See: http://mathworld.wolfram.com/BonferroniCorrection.html.

corrected, the critical value here is $p \le 0.025$. Accordingly, all targets with $p > 0.025$ were designated as being above the stop-digging threshold in our prioritized; otherwise, below.

Figure 53 shows the results of applying risk analysis to the blind data on this Combined-track. Again, we reduce the four-dimensional input set to two principal components for easy visualization (thus, this is what we have previously referred to as Attribute Set 3PC on the axes). The small, light-blue circles in Figure 53 are the blind data that appear above our stop-digging threshold while the darker blue circles appear below the stop-digging threshold. One can easily see the confidence boundary constructed around the UXO cluster (the red circles) by the residual risk analysis process.

Figure 53. Risk analysis boundary on Combined-track training and blind data



7.8 PRIORITIZED DIG-LIST PREPARATION

At this point, we had three sets of targets that needed to be combined into a single prioritized dig-list:

1. Cannot-Analyze Targets
2. Targets excluded as high-probability Not-UXO with the pre-discriminator;
3. The ranked targets from the LGP Discriminator.

In the experimental plan, cannot-analyze targets go at the bottom of the prioritized dig-list. Targets that are ranked by the LGP Discriminator as above the stop-digging threshold appear at the top of the list. Therefore, two sets of targets should appear below the stop-digging threshold:

1. Targets excluded as high-probability Not-UXO with the pre-discriminator; and
2. The targets ranked by the LGP Discriminator as below the stop-digging threshold.

These two sets of targets were combined using the $P(UXO)$ generated by our residual risk analysis. For item 1 the $P(UXO)$ used was the $P(UXO)$ generated by the pre-discriminator residual risk analysis that excluded the target. For targets described in 3, the $P(UXO)$ used was the value generated by the residual risk analysis on the LGP-generated scores.

7.9 DESCRIPTION OF IMPORTANT ATTRIBUTES IDENTIFIED BY LGP ON COMBINED-TRACK

Table 18 shows the ranking of Attribute Set 3 in the thirty bagging runs that produced our final LGP Models. The attributes are ranked by the frequency in which they appear in the best evolved programs across all bagging runs.

Table 18. Relative Importance of Attributes Used in LGP Modeling

Attribute Description	*LGP Frequency*	*Importance Evaluation*
EM_Attribute_3	0.961333333	Important
EM_Attribute_6	0.918333333	Important
EM_Attribute_1	0.595333333	Moderately Important
Mag_Attribute_2	0.335666667	Least Important

Two of the EM attributes were assessed as important and one as moderately important. On the other hand, the sole MAG attribute was relatively unimportant, appearing in only about 1/3 of the best programs across the forty bagging runs. The later fact is not too surprising, given our earlier observation that the sole MAG attribute did not appear to improve class separation when it was added to the attribute set (compare Figure 47 and Figure 48).

These attributes may be described as follows:

- EM_Attribute_3: The ratio of the signal values in the second decay channel to the signal values in the third decay channel in the centermost portion of the ellipse.
- EM_Attribute_6: The ratio of the signal values in the top coil to the signal values in the first decay channel across the entire ellipse.
- EM_Attribute_1: The variance of the ratio of the signal values in the second decay channel to the signal values in the third decay channel in the centermost part of the ellipse.
- MAG_Attribute_2: The distance between the high value in the positive lobe of the magnetic signal and the low value in the negative lobe of the magnetic signal.

7.10 FURTHER ITERATIONS

Because of the high quality of the results produced in the first iteration (reported above), the ESTCP Program Office suggested that we not perform any more iterations and we agreed with that conclusion on the ground that the classification portion of the ROC curve could not be improved in a statistically significant manner, even with more ground-truth.

8 DATA ANALYSIS AND PRODUCTS FOR INVERSION-TRACK

The Inversion-track used the phenomenological features created by inverting the MAGMTADS and EM61MTADS data for all targets selected by the program office as an MAGMTADS target, an EM61MTADS target, or both.

These phenomenological features were then used as basis for UXO discrimination by LGP.

The key difference between this track and the Combined-track is that in the Inversion-track, the derived phenomenological features are used as filters between the DGM and the LGP algorithm. By way of contrast, in the Combined-track, the LGP feature set is used as a filter between the DGM and the LGP algorithm.

The steps in this track were:

1. Combine the EM and MAG target sets;
2. Filter the EM and MAG targets to contain only targets where the phenomenological features are likely to contain useful information for discrimination;
3. Attribute extraction and reduction;
4. LGP modeling;
5. Residual Risk Analysis; and
6. Blind-scoring analysis

This section will first describe the combined EM61MTADS and MAGMTADS data and then summarize our process and results for each of those steps for the Combined EM/MAG track.

8.1 DESCRIPTION OF DATA

For this track, we extracted features for targets that were identified by the program office as targets for either the EM61MTADS sensor or the MAGMTADS sensor ("Inversion-track Targets"). Accordingly, there were more targets on this track than on the EM-only-track.

We received features for 1201 Inversion-track Targets. The 1201 targets are comprised of:

- 218 training (or "labeled") targets (targets for which we knew ground truth); and
- 983 blind data targets (targets for which we did not know ground truth).

Viewed another way, the 1201 Inversion-track Targets are comprised of:

- 712 targets that were selected by the program office as BOTH EM61MTADS targets and as MAGMTADS targets;
- 194 targets that were selected by the program office as EM61MTADS targets but not as MAGMTADS targets; and
- 295 targets that were selected by the Program Office as MAGMTADS targets but not as EM61MTADS targets.

8.2 ATTRIBUTE EXTRACTION

For this track, we extracted phenomenological attributes from the EM61MTADS signal and, separately, phenomenological attributes for the MAGMTADS signal. Those attributes are set forth in Table 19 and Table 20, together with a brief description of the information we received for each set of attributes. These tables also indicate whether the information was used in our further analysis.

Table 19. Summary of use of EM61MTADS inversion features

Name	*Description*	*Used in Further Analysis?*
TID	Target Identifier	No
X	Program Office Selected X	Limited
Y	Program Office Selected Y	Limited
EM_Fit_X	EM Inversion X Coordinate	Limited
EM_Fit_Y	EM Inversion Y Coordinate	Limited
EM_Fit_Depth	EM Inversion Depth	Yes
EM_Fit_Coh	EM Inversion Coherence. Measures fit of predicted signal to observed signal	Yes
EM_Fit_Size	EM Inversion Size	Yes
EM_Fit_Error	Flags an EM Inversion that did not converge	Limited
EM_Fit_b1	First polarization parameter	Yes
EM_Fit_b2	Second polarization parameter	Yes
EM_Fit_b3	Third polarization parameter	Yes
EM_Fit_theta		No
EM_Fit_phi		No
EM_Fit_psi		No
EM_Fit_chi2	Measures fit of predicted signal to observed signal	Yes

Table 20. Summary of use of MAGMTADS Inversion features

Name	*Description*	*Used in Further Analysis?*
TID	Target Identifier	No
Mag_X	Program Office Selected X	Limited
Mag_Y	Program Office Selected Y	Limited
Mag_Fit_X	Mag Inversion X Coordinate	Limited
Mag_Fit_Y	Mag Inversion Y Coordinate	Limited
Mag_Fit_Depth	Mag Inversion Depth	Yes
Mag_Fit_Coh	Mag Inversion Coherence. Measures fit of predicted signal to observed signal	Yes
Mag_Fit_Size	Mag Inversion Size	Yes
Mag_Fit_Error	Flags a Mag Inversion did not converge	Limited
Mag_Fit_Dec		No
Mag_Fit_Inc		No
Mac_Fit_Solid_Angle		Yes
Mag_Fit_MagMoment		Yes

Some of the features described in Table 19 and Table 20 were either not used at all or used in a limited role in further analysis as follows:

1. The X, Y, Fit_X and Fit_Y values were used to only to compute the distance between the Program Office X,Y coordinates and the coordinates produced by the inversion ("Fit_Distance"). The EM and Mag Fit_Distances were used only for assessment of assigning a "cannot-analyze" label to targets where the inversion moved the fit location an implausible distance.
2. EM_Fit_Error and Mag_Fit_Error were used only to exclude targets from further analysis as "cannot-analyze" targets.
3. EM_Fit_Theta, EM_Fit_Psi and EM_Fit_Pi were excluded from further analysis because insufficient theoretical or empirical evidence exists to support their inclusion as a proper discriminator.
4. Mag_Fit_Dec and Mag_Fit_Inc were excluded because they contain the same information as Mag_Fit_Solid_Angle. Accordingly, we chose the more parsimonious attribute with which to continue.

8.3 CANNOT-ANALYZE FOR THE INVERSION-TRACK

This section describes the issues raised by the Inversion-track attributes in terms of assigning targets to the cannot-analyze category. Of course, our goal in marking cannot-analyze targets was to exclude targets where the attribute set was not sufficiently reliable on which to base a classification while keeping the number of cannot-analyze targets to a minimum.

It became clear when we analyzed the data for this track that achieving these goals would be difficult because of the number of targets for which there was an obvious problem with at least one of the attributes extracted.

8.3.1 EM and Mag Coherence Data Quality Issues on Inversion-track

For our first pass at this issue, we tried an industry standard cutoff for both Mag and EM Coherence of 0.95. That is, any target with either Mag or EM Coherence of less than 0.95 was labeled cannot-analyze. While that produced very good discrimination results, 66% of the Blind targets had EM_Fit_Coherence < 0.95 while 55% of the blind targets had a Mag_Fit_Coherence < 0.95. Altogether, using this criterion for cannot-analyze, we would have been required to exclude 77% of all blind targets as cannot-analyze targets. That would turn this track into an interesting academic exercise with no practical application.

Accordingly, after discussion amongst the P.I.'s we first decided to use an EM_Fit_Coherence threshold of 0.85 and a Mag_Fit_Coherence threshold of 0.65. For discrimination purposes, it is preferable to have all attributes (Mag and EM) consistent and defensible. However, if we required both EM and Mag coherence measures to meet the above criteria, 52% of the blind targets would have to be excluded as cannot-analyze by that measure alone (See Table 21). Again, this is an unacceptably high number. (Even by this criterion, the cannot-analyze blind targets would be higher than 52% of all blind targets because there were other problems with the inversion features described in Table 21).

Table 21. Summary of cannot-analyze issues and effected targets for Inversion-track

Issue Type	*Issue Criterion*	*Percent of Blind Targets Effected*	*Percent of Train Targets Effected*
Fit Error (Mag)	Occurred	14%	16%
Fit Error (EM)	Occurred	3%	3%
Low Coherence (Mag OR EM)	EM_Coh < 0.85 OR Mag_Coh < 0.65	52%	46%
Low Coherence (Mag AND EM)	EM_Coh < 0.85 AND Mag_Coh < 0.65	13%	13%
Implausible Depth	Mag_Fit_Depth > 3 Meters or EM_Fit_Depth > 2 Meters or either Fit_Depth feature describes an object suspended in the air	12%	8%
Implausible distance moved during inversion	> 1 Meter	21%	16%

After discussion amongst the PI's, we decided to limit the coherence cannot-analyze criterion to:

$$(EM\_Fit\_Coherence < 0.85) \cap (Mag\_Fit\_Coherence < 0.65),$$

where $\cap$ is the logical AND operator. The key determination here was that we required only one of the two inversions to show coherence that exceeded the relaxed thresholds. In other words, we

would model targets where ONLY ONE of the two inversions (EM or Mag) meets a minimal inversion threshold.

This new criterion, by itself, would require that we assign 13% of the blind data targets to the cannot-analyze category (See Table 21). This is a great improvement over the 52% cannot-analyze when we required BOTH of the inversions to meet minimal coherence thresholds.

8.3.2 Additional Data Quality Issues on Inversion-track

Low coherence is only the most significant problem with the inversion data. There are others, also described in Table 21 that effected our cannot-analyze criteria. Specifically:

- A Mag_Fit_Error or an EM Fit_Error indicates the inversion did not converge.
- Many inversions produced a depth feature that is implausible either because it is too deep, given the detection capabilities of the sensor, or because it describes a metal object suspended in the air.
- Finally, many inversions produce location features that are more than one meter from the target location picked by the Program Office.

Collectively, these three exclusion criteria, together with the relaxed coherence criterion we adopted, would result marking approximately 44% of the blind targets as cannot-analyze targets. Again, this seems unacceptably high. The next section describes the process we devised to reduce the number of cannot-analyze targets down to 26% of the blind targets in a statistically defensible manner.

8.3.3 Reducing the Number of Cannot-Analyze Targets for the Inversion-Track using EM_Fit_Coherence and EM_Fit_Size Based Pre-Discriminators

The problem on this track, like the previous two tracks, was really that, between the number of probably-not-metal targets identified by the MAG sensors and the rut noise that affects the EM sensors, the inversions fail to produce proper inversions or even minimal coherence figures because, in all likelihood, there is nothing there.

We successfully tested and implemented two sequential pre-discriminators to filter a portion of these targets. This is analogous to our “amplitude discriminator” used for the other two tracks.

Our goal here was to find one or more simple discriminators with which we can exclude as many targets as high probability Not-UXO as possible, without excluding the low coherence, implausible depth etc targets. The goal is not to do a sophisticated multi-dimensional model that will discriminate the difficult Not-UXO (e.g. half-shells) from the UXO. The goal is to pick off as many of the easy high-probability Not-UXO as possible so as to reduce the number of cannot-analyze targets in a statistically proper manner.

We limited our analysis of attributes to the EM attributes for this step because the EM inversions produced far fewer fit errors (3% of blind targets) than did the MAG inversions (14% of blind targets).

To proceed with this analysis we first assigned all targets that produced an EM_Fit_Error as cannot-analyze targets. Altogether, 38 targets were assigned to cannot-analyze. After that, we had 951 blind targets and 212 training targets remaining.

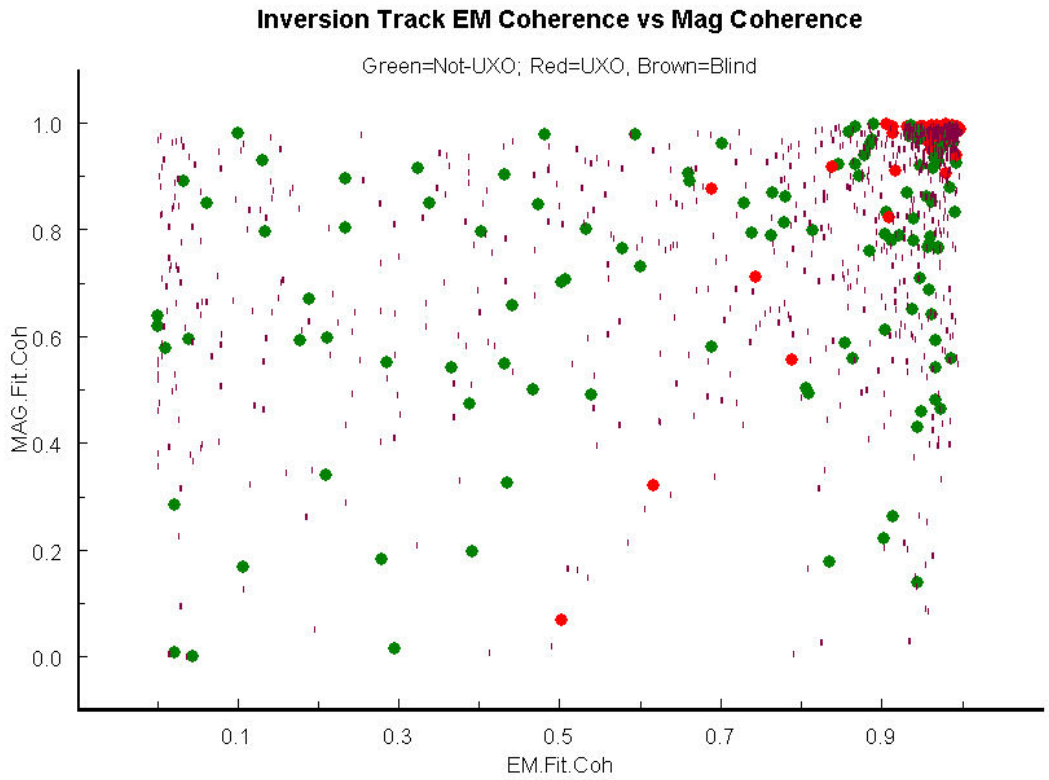
We located two attributes that appeared to do a good job in identifying reasonably large groups of small targets, EM_Fit_Coherence and EM_Fit_Size, in that order. What follows is our analysis and discriminator derivation using each of them.

8.3.3.1 EM_Fit_Coherence Discriminator

We used EM_Fit_Coherence as our first filter to reduce the number of cannot-analyze targets. Our process for doing so was in four steps: (1) Visual inspection of the data; (2) Assessment of the statistical significance of the feature for excluding Not-UXO as high-probability Not-UXO; and (3) Visual comparison of the distribution on this feature of the training and blind data to assure that the training data is reasonably representative of the blind data on this feature; and (4) Residual Risk Analysis using the feature.

We started by checking our decision, noted above, to use only EM features for the pre-discriminators. Figure 54 permits a visual comparison of EM_Fit_Coherence and MAG_Fit_Coherence as a discriminator.

Figure 54. EM_Fit_Coherence vs. MAG_Fit_Coherence as a Discriminator.



The green circles in Figure 54 are labeled data that is Not-UXO. The red are UXO. The small brown dots are blind data. MAG_Fit_Coherence (the y-axis) is obviously a poor discriminator for the goals of this preliminary filter. We could not safely exclude any targets as high-probability Not-UXO based on MAG coherence. On the other hand, the EM_Fit_Coherence feature concentrates all training UXO in the values of 0.5 and greater.

Furthermore, the EM_Fit_Coherence provides a highly statistically significant split of the training data into UXO and Not-UXO. The UXO with the lowest EM_Fit_Coherence has

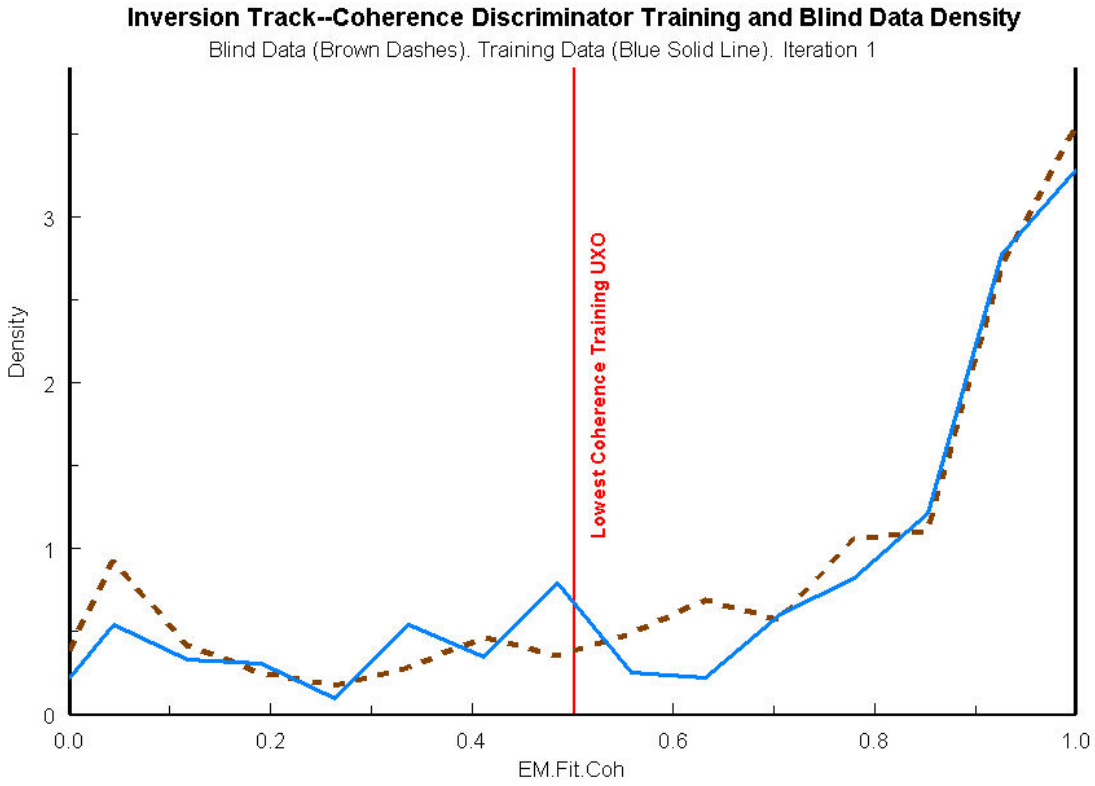
EM_Fit_Coherence of 0.52. If we split the training data at EM Coherence=0.5, we obtain the following 2x2 contingency table for UXO and Not-UXO above and below the split:

Table 22. Two-by-Two contingency table for EM_Fit_Coherence as a UXO discriminator

	Below Split	Above Split
UXO	0	59
Not-UXO	43	41

The Chi Squared statistic for the relationship shown in this table, corrected for continuity, is 40.79 with one degree of freedom. The probability of Chi Square for this table is 0.000 (in other words, zero to available machine precision). Accordingly, we conclude that the split of the training data at 0.5 using EM_Fit_Coherence produces a highly statistically significant separation of Not-UXO from other targets. Given this significant separation of Not-UXO from other targets and given the good match between the densities of the training and blind data, we selected EM_Fit_Coherence as our first pre-discriminator.

Figure 55. Comparative Density of Blind and Training Data on EM Coherence



Next, Figure 55 shows there is an excellent match between the density of training and blind data using EM_Fit_Coherence. So we expect the training data to provide robust results for the blind data.

The next task in this process is to determine where, on the EM_Fit_Coherence axis, we may safely say that the probability that all items with lower EM_Fit_Coherence are Not-UXO. To do that, we turn to our residual risk analysis methodology.

We first converted the EM_Fit_Coherence values into ranks across the entire training and blind data sets. In making this conversion, lower values of EM_Fit_Coherence were interpreted as higher rankings. We then evaluated logistic regression, exponential regression, power law regression and kernel regression as tools to fit the probability of UXO as a function of rank.

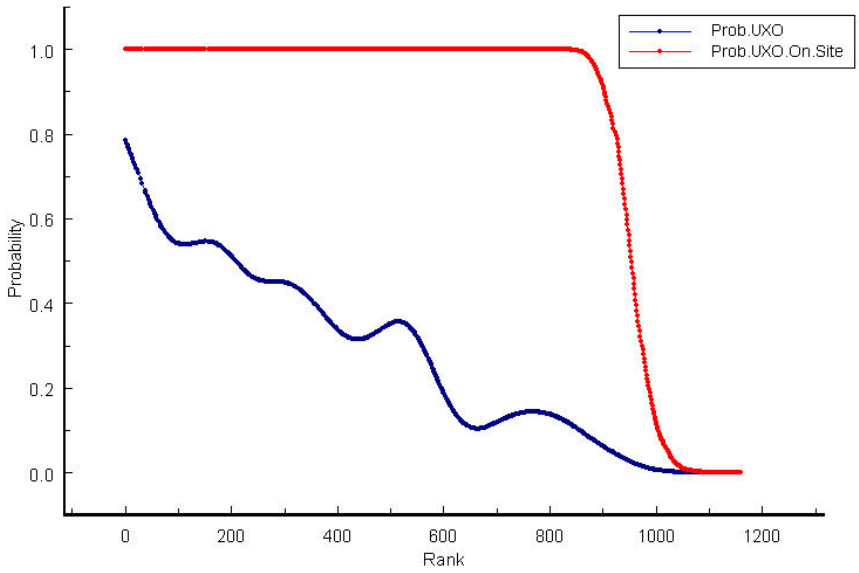
The first three functional types were deemed inappropriate because of the local ups and downs of the probability as a function of Amplitude Principal Component 1 (see Figure 56). Kernel regression, on the other hand, does a good job of modeling such local irregularities and is generally preferable to the others, all other things being equal, because it is a single-parameter model. Accordingly, we used kernel regression with a Gaussian kernel as set forth in Equation 3 to model probability of UXO.

We derived the width parameter, α, for Equation 3 using leave-one-out cross-validation on the training data, optimizing the value of the parameter in the manner as described in Section 6.8.4. The value determined for the parameter, α, is 53.084.

Next, we applied the Gaussian kernel, generated by the training data, using the derived kernel width α parameter, to the ranked blind data. This generated a probability that each blind data target is UXO. Figure 56 shows that probability as a function of rank (blue series) on the blind targets.

Once those probabilities were predicted on the blind targets, we then assessed the probability that the blind targets ranked above each Amplitude Principal Component 1 ranking contain one-or-more UXO. To do so, we used the "or-of-probabilities" approach described in Section 2.1.6, Equation 2, applied to all such higher ranked targets. This generates the cumulative probability that one-or-more UXO remain on site above each ranking. Figure 56 shows that cumulative probability as a function of rank (red series) on the blind targets.

Figure 56. Probability of UXO and probability of UXO remaining on site as a function of EM_Fit_Coherence rank. Blind targets.



When the red series in Figure 56 falls below a critical p value, we assess all targets remaining to the right of that value as high-probability Not-UXO.

The critical value we used was the Bonferonni corrected p-value for a 95% confidence level. We use the corrected value because we are using three discriminators on this track.[31] The critical value here is $p \le 0.01667$. Using that criterion, we select $EM\_Fit\_Coherence \le 0.127$ as the point below which we will assign targets to high-probability not-MEC. At that point, the probability of remaining UXO is 0.0158—in other words, it satisfies the $p \le 0.01667$ criterion, above.

The result of this process excludes 105 blind targets and 14 training targets as high probability Not-UXO. This step was modestly successful because many of those blind targets would have had to be excluded as cannot-analyze targets if we were required to utilize other inversion features for discrimination.

After applying this EM_Fit_Coherence discriminator, there were 846 blind and 198 training targets remaining for analysis.

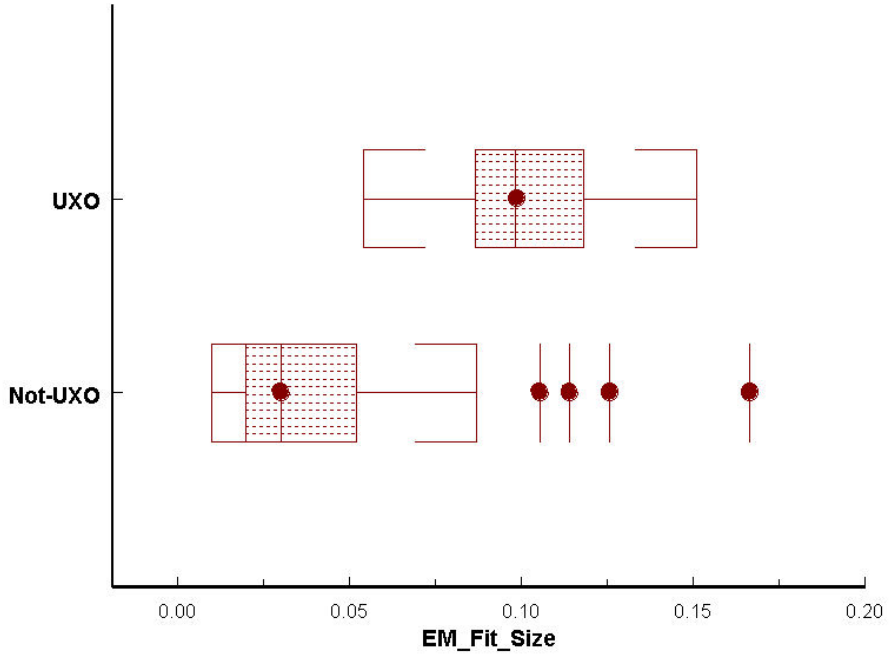
8.3.3.2 EM_Fit_Size Discriminator

We used EM_Fit_Size as our second filter for to reduce the number of cannot-analyze targets. Our process for doing so was in four steps: (1) Visual inspection of the data; (2) Assessment of the statistical significance of the feature for excluding Not-UXO as high-probability Not-UXO; and (3) Visual comparison of the distribution on this feature of the training and blind data to assure that the training data is reasonably representative of the blind data on this feature; and (4) Residual Risk Analysis using the feature.

[31] See: http://mathworld.wolfram.com/BonferroniCorrection.html.

To begin with, we assess EM_Fit_Size visually. Figure 57 shows the distribution of UXO and Not-UXO on the EM_Fit_Size feature. This appears to be a good discriminator to eliminate smaller targets because at least 75% of the EM_Fit_Size values of the Not-UXO (the high end of the lower shaded box) are less than the minimum UXO value for EM_Fit_Size (0.054).

Figure 57. Distribution of UXO vs. Not-UXO on EM_Fit_Size feature. Training data only.



Next, we check the statistical significance of splitting the data using EM_Fit_Size. The UXO with the lowest EM_Fit_Size has EM_Fit_Size of 0.054. If we split the training data at EM_Fit_Size <0.054, we obtain the following 2x2 contingency table for UXO and Not-UXO above and below the split:

Figure 58. Two-by-two contingency table for splitting UXO from Not-UXO using EM_Fit_Size

	Below Split	Above Split
UXO	0	59
Not-UXO	106	33

The Chi Squared statistic for this table is 43.2 with one degree of freedom. The probability of Chi Squared is 0.000. Accordingly, we conclude that the split of the training data at EM_Fit_Size< 0.054 produces a highly statistically significant separation of Not-UXO from other targets.

Next, we check that the distribution of the training and blind data for EM_Fit_Size is reasonably matched. Figure 59 and Figure 60 show histograms of the two distributions with a density plot overlay. The match is generally reasonable. However, the training data does have a long tail to the right that is considerably more substantial than the blind data. This is somewhat odd because

the blind data is much larger than the training data. So we would expect the blind data to have longer tails in the distribution on both ends. For this reason, we will be conservative when selecting functional fits for the residual risk analysis, as described below.

Figure 59. Histogram and density plot of training data for EM_Fit_Size

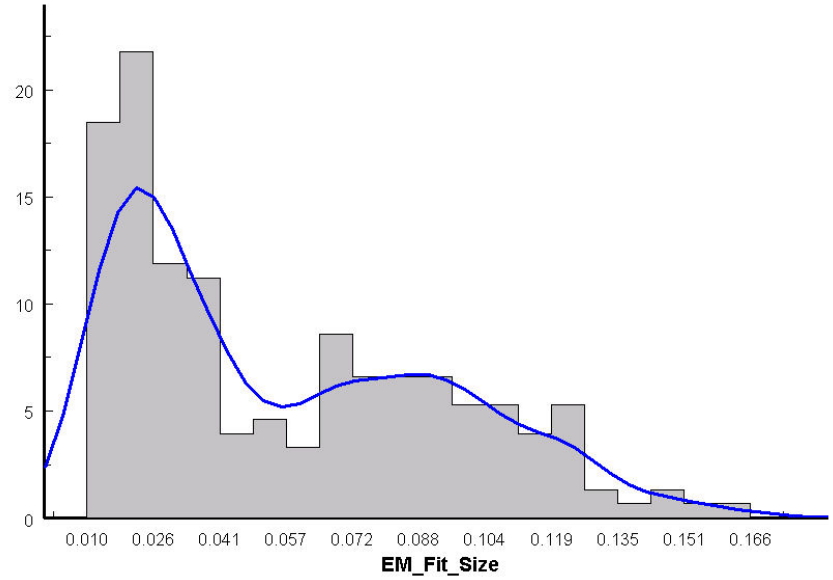
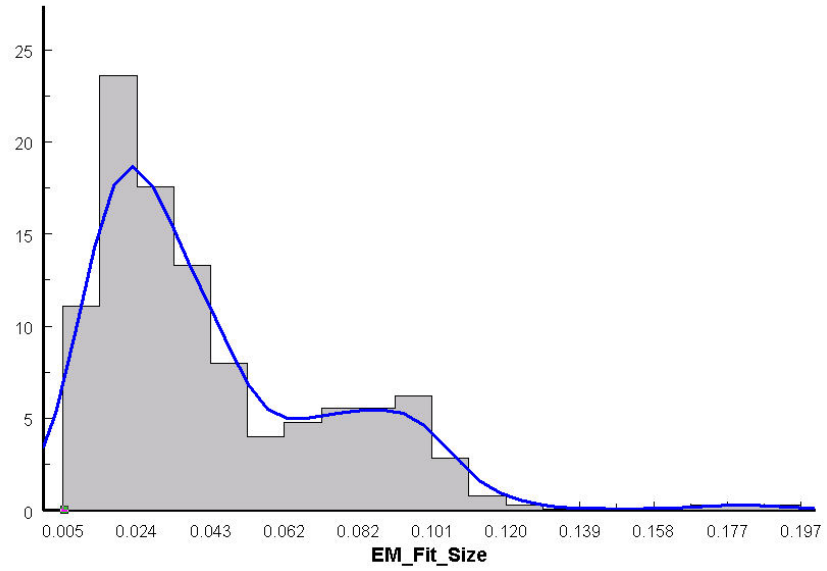


Figure 60. Histogram and density plot of blind data for EM_Fit_Size



The next task is to determine where, on the EM_Fit_Size axis, we can safely say that the probability that all items with lower EM_Fit_Size are not-UXO. To do that, we turn to our risk analysis methodology.

We first converted the EM_Fit_Size values into ranks across the entire training and blind data sets. In making this conversion, lower values of EM_Fit_Size were interpreted as higher rankings.

We then evaluated logistic regression, exponential regression, power law regression and Gaussian kernel regression to model probability of UXO as a function of EM_Fit_Size rank. Ordinarily, we would use kernel regression for this process as discussed above. However, kernel regression is often more aggressive than logistic regression on these data in excluding blind data as high-probability Not-UXO. Because of the mismatch in the tail of the training and blind data on this variable, we made a judgment call to use the more conservative measure of logistic regression to model the falling probability of UXO as a function of EM_Fit_Size rank using the logistic transform.

Accordingly, we performed logistic regression on the training data for the current step, which optimizes two parameters in the functional form shown in Equation 4. The dependent variable was the groundtruth labels on the training targets and the independent variable was the EM_Fit_Size based ranks.

We derived values for the two parameters in Equation 4 using leave-one-out cross-validation and standard logistic regression. The parameter values were:

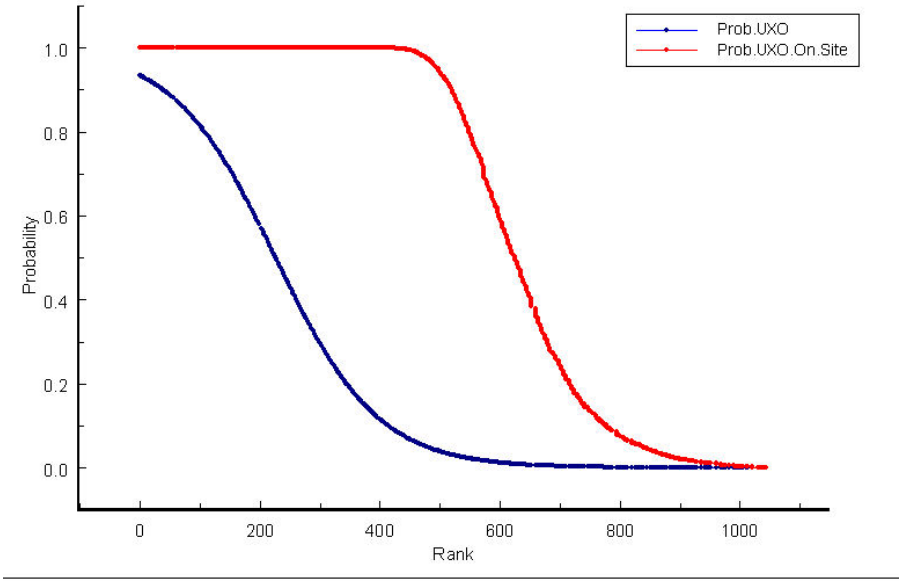
$$\alpha = 2.6457$$

$$\beta = -0.01171$$

Then, we substituted these parameter values into Equation 5 to predict the probabilities of UXO on the blind data by rank, using the ranks derived from EM_Fit_Size as the independent variable. These probabilities are shown in the blue series in Figure 61 for the blind targets remaining at this point in the Inversion-track.

Once we derived these probabilities for each blind target, we calculated for each rank, the cumulative probability that one-or-more of the blind targets that have a higher ranking than the rank for which we are making the calculation contain UXO. Those cumulative probabilities are calculated using the “or-of-probabilities” approach described in Equation 2 in Section 2.1.6. These cumulative probabilities that UXO remains on the site are shown in the red line in Figure 61 for the blind targets remaining at this point in the Inversion-track.

Figure 61. Residual risk analysis for EM_Fit_Size as a high-probability not-UXO discriminator



When the red series reaches a critical p-value, we assess all targets remaining to the right of that ranking as high-probability Not-UXO.

Because we used three discriminators here (the EM_Fit_Coherence filter, the EM_Fit_Size filter and the LGP Discriminator), each of which is subjected to probabilistic risk assessment, we used the Bonferonni corrected p-value for a 95% confidence level of p=0.016667. Using that criterion, we selected EM_Fit_Size<= 0.0159 as the cutoff. At that ranking corresponding to that value of EM_Fit_Size, the probability of remaining UXO is 0.0164.

The result of this process excludes 124 targets in total as high-probability Not-UXO, including 107 blind targets and 17 training targets.

After applying this EM_Fit_Size discriminator, there were a total of 920 targets remaining for analysis, including 739 blind and 181 training targets.

8.3.4 Exclude Cannot-Analyze Targets Remaining after Pre-Discriminators

At this point, we had done about as much as possible to reduce the potential set of cannot-analyze targets for the Inversion-track. Accordingly, we applied the following criteria to exclude targets as cannot-analyze. The key idea behind each of these criteria was to assure that the inversion provided probably valid results for both MAG and EM inversions from which LGP could build valid models.

Note that these criteria are applied in order. As a practical matter, many of the targets excluded as cannot-analyze would have been excluded under multiple criteria. However, by applying the criteria sequentially, each target is counted only once in the following list.

8.3.4.1 MAG_Fit_Error Targets

As noted above, we started with 179 targets for which the Mag inversion did not converge. We were able to assign 75 of those targets to High-Probability Not-UXO using the two EM-based

statistical discriminators discussed above. The remaining 130 Mag_Fit_Error targets, including 27 training targets and 103 blind targets were assigned to cannot-analyze.

8.3.4.2 Low EM_Fit_Coherence and Low MAG_Fit_Coherence

After considerable discussions amongst the P.I.'s, it was determined to use the following criteria to exclude targets based on their coherence numbers. We excluded only targets that had both low EM_Fit_Coherence and MAG_Fit_Coherence. The criterion used was:

$$\left(EM\_Fit\_Coherence < 0.85\right) \cap \left(MAG\_Fit\_Coherence < 0.65\right)$$

Where $\cap$ represents the logical AND operator.

Altogether, 86 targets that met this criterion, including 16 training and 70 blind targets. These targets were assigned to cannot-analyze.

8.3.4.3 Implausible EM_Fit_Depth or Implausible MAG_Fit_Depth

Another sign that an inversion has produced an invalid result is if it generates a fit depth that is improbable, given the equipment and targets at issue. We did NOT exclude targets that had depth figures for either MAG or EM inversions using the following criterion:

$$\left(-0.1 < EM\_Fit\_Depth < 2\right) \cap \left(-0.1 < MAG\_Fit\_Depth < 3\right)$$

Where $\cap$ is the set AND operator and the depths are in meters.

In our inversions, a negative depth indicates an object above the surface. By allowing a margin of 10 centimeters above the surface, we allow for the possibility of small objects on the surface but exclude inversions that suggest the metallic object is hovering in the air. The depth thresholds on the low end are set to values lower than the lowest values at which we expect to be able to detect the target ordnance, given the sensor set.

Altogether, 16 targets did not fall within an acceptable depth range, including 2 training and 14 blind targets.

8.3.4.4 MAGXY_Dist_Moved or EMXY_Dist_Moved

We also used the distance the inversion moved the x,y coordinates of the targets as a metric for identifying probably invalid inversions. The criterion used was:

$$\left(EMXY\_Dist\_Moved > 1\right) \cup \left(MAGXY\_Dist\_Moved > 1\right)$$

Where $\cup$ is the set OR is operator and the constants are measured in meters.

Altogether, 32 targets met this criterion, including 7 training and 25 blind targets. These targets were assigned to cannot-analyze.

8.3.5 Check for Remaining Outliers on Polarization Parameters

We checked whether the targets remaining after the above process produced credible inversions on the polarization parameters for the remaining targets. To do so, we normalized the three EM polarization parameters by converting them to z-scores and then identified outliers amongst all remaining training and blind targets on the three normalized z-scores. We used a robust Mahalanobis distance to identify outliers at the 99% confidence level. We then examined which

targets amongst the training targets had been identified as outliers. Altogether 60 training targets were identified as outliers. Of those, 50 were UXO, 5 were Half Shells and one was a fragment item.

This is, of course the result we would expect if the remaining targets were producing mostly credible inversions. UXO *should* stand out from the other targets.

We note here that three of the outliers were identified in the groundtruth as "Soils". We would not expect soil to produce a valid inversion because no metal was located. This probably means that a portion of the inversions we have not excluded as cannot-analyze targets did not produce good inversions. All three of these soils targets had MAG Coherence above the 0.65 threshold but EM Coherence below 0.70.

It would be tempting to change the coherence threshold to remove these "soils" targets. However, the cost of excluding all "soils" targets by changing the coherence criterion for "cannot-analyze" would be greatly to increase the number of cannot-analyze targets, including many substantial metal targets, as discussed above. In addition, these three targets all produced credible inversion parameters.

Accordingly, we elected not to change the coherence criterion so as to identify these soils targets as cannot-analyze target and to leave the task of distinguishing the remaining soils targets from legitimate targets to the LGP discriminator. We regarded this as an acceptable risk because the error that would be expected from a "Soils" target that looks like a UXO is a false positive, not a false negative. That is, an error of this type would not result in leaving UXO in the ground.

8.3.6 Conclusions Regarding Pre-Discriminators and Cannot-Analyze Targets

The point of the two pre-discriminators on this Inversion-track was to reduce the very high proportion of cannot-analyze targets produced by preliminary analysis of the inversion features. Thus, instead of classifying all of these targets as cannot-analyze, we were instead able to assign many of them to high-probability not-UXO. At this point, it was possible to assess the effect of the two pre-discriminators on the number of cannot-analyze.

Before applying the two pre-discriminators, 44% of the blind data would have been classified as cannot-analyze using the criteria outlined above. Using the same cannot-analyze criteria, after applying the two pre-discriminators, only 26% of the blind data had to be classified as cannot-analyze. While this is still not nearly as good as the EM-only-track and the Combined-track results, it is nevertheless a significant improvement in the performance on this Inversion-track.

Of course, the targets that have been excluded to this point, either as high-probability Not-UXO or as cannot-analyze play no part in the next several steps in our process. In particular, they play no role in the attribute reduction step, the LGP modeling step, or the residual risk analysis step.

8.4 ATTRIBUTE REDUCTION

Having removed the cannot-analyze targets and the high-probability non-UXO identified by the pre-discriminators, we then proceed to the attribute reduction step. On the Inversion-track, attribute reduction was simple and proceeded in two steps: (1) We combined certain highly correlated attributes using principal components; and (2) We removed one attribute based on a combination of Mutual Information ranking and visual inspection of attribute space.

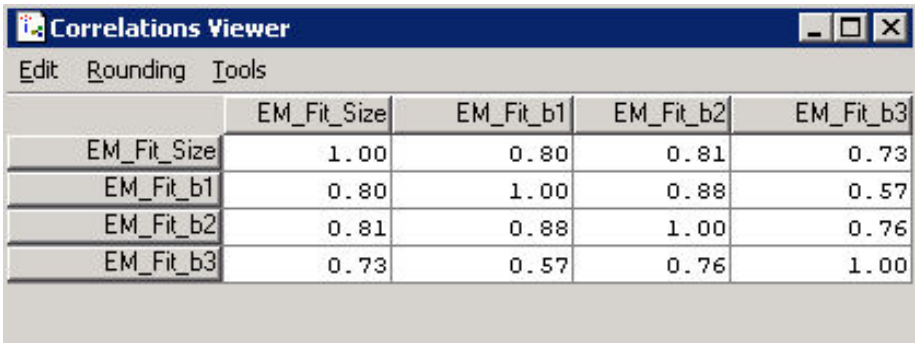
Table 19 and Table 20 show our starting point in this analysis. All features in those tables that are marked “Yes” in the “Used in Further Analysis” columns comprised the starting point for our attribute reduction process on the Inversion-track.

8.4.1 Replace Highly Correlated EM Features with Principal Components

We examined a correlation matrix for the features marked in Table 19 and Table 20 as “Used in Further Analysis.” It was immediately obvious that four EM features were highly correlated amongst themselves. They were:

1. EM_Fit_Size
2. EM_Fit_b1
3. EM_Fit_b2
4. EM_Fit_b3

Figure 62. Correlation matrix for four highly correlated EM features



	EM_Fit_Size	EM_Fit_b1	EM_Fit_b2	EM_Fit_b3
EM_Fit_Size	1.00	0.80	0.81	0.73
EM_Fit_b1	0.80	1.00	0.88	0.57
EM_Fit_b2	0.81	0.88	1.00	0.76
EM_Fit_b3	0.73	0.57	0.76	1.00

Figure 62 shows the correlation coefficients for these four features. Using principal components analysis, it is simple to reduce these four features to two features. We will refer to these components as the “EM_Size” group or correlation cluster components.

Figure 63. EM_Size group principal components

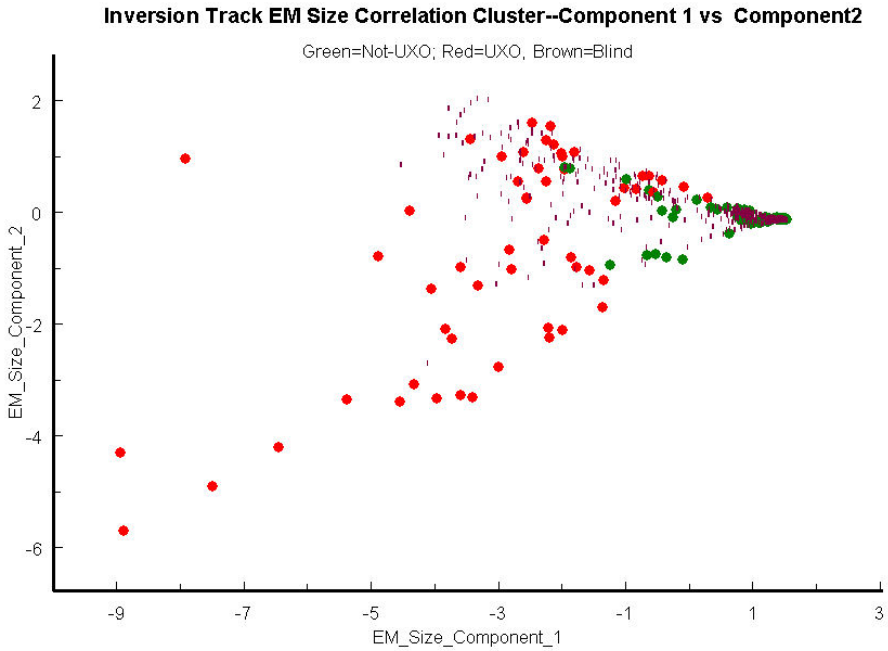


Figure 63 shows the training and blind data on the two EM_Size group principal components. As usual, UXO is red, not-UXO is green, and blind data are the small brown dots. Both components significantly split the UXO from not-UXO. Accordingly, we will use the two components in place of the four highly correlated EM Size features.

8.4.2 Remove Features Based on Visual Inspection of Attribute Space

The next feature reduction step we took was to rank the features (including the principal components) using a mutual information criterion that takes into account mutual information between the features and the ground truth and also the redundancy of mutual information amongst the features themselves. To do so, we binned each of the potential features into eight bins and then ranked them by the MRMR criterion. Table 23 shows the result of that process.

Table 23. Ranking of inversion features for potential predictive power

Rank	Column Name	Mutual Information With Groundtruth	Mutual Information with Previously Ranked Features
0	EM_SIZE_COMPONENT_1	0.707958192	2.910185
1	MAG_FIT_SOLID_ANGLE	0.155827836	0.354493
2	MAG_FIT_MAGMOMENT	0.445405022	0.463226
3	MAG_FIT_COH	0.329952098	0.452394
4	MAG_FIT_DEPTH	0.29574388	0.464618
5	EM_FIT_COH	0.117126436	0.35204
6	EM_SIZE_COMPONENT_2	0.490703798	0.702689
7	EM_FIT_DEPTH	0.205602418	0.504084

8	MAG_FIT_SIZE	0.496447042	0.789885
9	EM_FIT_CHI2	0.31636607	0.614198

To see if any further feature reduction was warranted, we visually examined the bottom three ranked features graphed against EM Size Component 1. Of them, EM_Fit_CHI2 appeared to contain little useful information in addition to the information contained in EM_Size_Component_1. Accordingly, EM_Fit_CHI2 was eliminated from the feature set and LGP was run on the remaining features shown in Table 23.

8.5 REMOVE FEATURE SPACE OUTLIERS AS CANNOT-ANALYZE

The final step before LGP modeling is to visually examine the feature space of the reduced features for outliers. Outliers are assigned to cannot-analyze. Table 24 shows the 16 targets excluded as cannot-analyze targets because they are attribute-space outliers.

Table 24. Feature Space Outliers Excluded as Cannot-Analyze Targets

Target ID	*Exclusion Reason*
1280	Outlier on EM_Fit_Component_1. EM_Fit_Component_1 < -7.7
1130	Outlier on EM_Fit_Size vs. Mag_Fit_Moment and on EM_Fit_Component_1 vs. Mag Fit Size and on EM_Fit_Component_1 vs. Mag_Fit_Solid_Angle. EM_Fit_Size > 0.18.
1137	Outlier on EM_Fit_Component_1 vs. Mag_Fit_Depth. Mag_Fit_Depth > 1.9
782	Outlier on EM_Fit_Component_2. EM_Fit_Component_2 > 2
1171	Outlier on EM_Fit_Component_2. EM_Fit_Component_2 > 2
1138	Outlier on EM_Fit_Component_1 vs. Mag_Coh
998	Outlier on EM_Fit_Component_1 vs. Mag_Coh
320	Outlier on EM_Fit_Component_1 vs. Mag_Coh
722	Outlier on EM_Fit_Component_1 vs. Mag_Coh
1269	Outlier on EM_Fit_Component_1 vs. Mag_Coh
1258	Outlier on EM_Fit_Component_1 vs. Mag_Fit_Size and on EM_Fit_Component_1 vs. Mag_Fit_MagMoment
874	Outlier on EM_Fit_Component_1 vs. Mag_Fit_SolidAngle
315	Outlier on EM_Fit_Component_1 vs. Mag_Fit_SolidAngle
1057	Outlier on EM_Fit_Component_1 vs. Mag_Fit_SolidAngle
528	Outlier on EM_Fit_Component_1 vs. Mag_Fit_SolidAngle
1056	Outlier on EM_Fit_Component_1 vs. Mag_MagMoment

Once these 16 targets were excluded, we then passed the reduced feature and target set to LGP classification, as described in the next section.

8.6 LGP DISCRIMINATION ON INVERSION-TRACK

LGP Discrimination used the above described feature set and took place in two steps: (1) Cross-validation to set the noise parameter; and (2) Bagging to produce a model and prioritized dig-list.

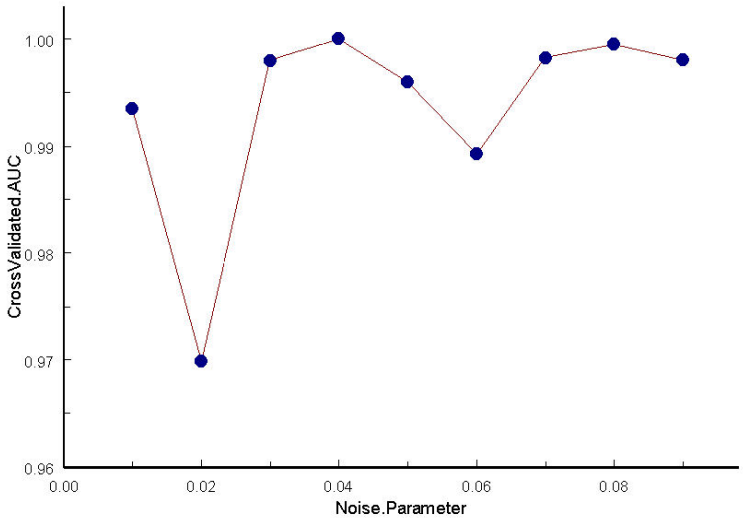
8.6.1 Cross-Validation to Set the Noise Parameter

This is a small training set. To prevent over-fitting the training data, we added a small amount of Gaussian noise to the inputs. The standard deviation of the added noise is set attribute by attribute. A noise parameter of 2% means that the standard deviation of the Gaussian noise is set to 2 percentiles of the distribution of that variable.

Setting the amount of noise is an empirical process dependent on the data set at hand. We set the noise parameter using ten-fold cross validation, testing noise settings of 1% thru 9% in increments of one. In performing the cross-validation, the default settings of Discipulus™ LGP software were used with the following exceptions: (1) The fitness function used was Area under the curve; (2) The termination criterion for each run was 40 generations without improvement; (3) The number of runs performed in each project was 20 runs. Of course, the noise level was varied for parameter selection.

Most noise settings produced an area under the ROC curve (AUC) summed over the held-out cross-validation data of 0.99 or better (a very good ROC curve). Figure 50 shows the cross-validated AUC over all tested noise settings. The two best noise parameter settings were 8% (AUC=0.9995) and 4%, (AUC=1) and these settings are statistically indistinguishable from each other. Accordingly, we selected the 4% and 8% noise settings for further analysis.

Figure 64. Cross-validated area under the curve for various noise parameter settings



8.6.2 Bagging to Produce the LGP Ensemble Model

To prepare the prioritized dig-list, we performed 40 bagging runs at each of the two selected noise parameter settings. The training data for each "bag" is selected by taking n samples (each sample being a specific training target together with all attributes and labels associated with that target) with replacement from the full training data set, where n is equal to the number of

training data points. The training targets NOT selected for that "bag" (about 32% of the training data in each "bag") are not used in training for that "bag". Rather, they are held-out from training process. These "held-out" training targets are referred to as the "out-of-bag" data.

The default settings of Discipulus™ LGP software were used with the following exceptions: (1) The fitness function used was Area under the curve; (2) The termination criterion for each run was 40 generations without improvement; (3) The number of runs performed in each project was 20 runs. Forty projects were run at the 4% noise level and forty projects were run at the 8% noise level. Each project used a different random "bag" for the training data.

Our final model is, therefore, an ensemble of eighty LGP evolved programs—forty at a 4% noise and forty at an 8% noise. Those eighty programs are referred to as an "LGP ensemble predictor."

8.6.3 Out-of-Bag Error to Estimate Performance on Blind Data

Predictions on the out-of-bag data are used to predict the expected error on the blind data and for residual risk analysis. They are used because the labels on the out-of-bag data are unknown to the LGP algorithm when it is training. Thus, the out-of-bag error is our best estimate of the expected error (1-AUC) on blind data.

We computed the out-of-bag error as follows: Each training target has multiple predictions from the LGP ensemble predictor that are produced when that target was in the out-of-bag data. Those predictions are summed for each training target and averaged. This average was treated as our prediction for that data point. The predictions, of course, permit us to rank the training data points relative to each other in a prioritized dig-list. That list produces a ROC Chart.

The out-of-bag ROC chart on this track is easy to summarize. All of the UXO are ranked above all of the not-UXO. Accordingly, the AUC on the out-of-bag training data is 1 and the expected error (1-AUC) is zero. We expect similar numbers for the blind data.

8.6.4 Scoring the Blind Data with LGP Models

We then score the blind targets using the same LGP ensemble predictor. The score for each blind target was the average of all outputs from the models in the ensemble for that target.

8.7 RESIDUAL RISK ANALYSIS FOR LGP MODELED TARGETS

This section describes the application of our risk analysis methodology to the LGP ensemble predictor described in the previous section for the Inversion-track.

In summary, we took the scores of the LGP ensemble predictor for both training and blind data for this step and assembled them to produce a combined ranking across both data sets. In making that conversion from scores to ranks, a low LGP score was converted to a high ranking (that is, a low LGP score translates to a ranking that is less likely to be UXO). Then, we built a regression model of the probability of UXO as a function of that rank, using that rank and the known groundtruth for the training data. Finally, we applied that regression model to the blind data and calculated the residual risk from the resulting probabilities for the blind targets

After assembling the ranks across all training and blind data for this track, the next step in risk analysis was to build a probabilistic regression model of the UXO/Not-UXO groundtruth as a function of the rank across the training and blind data in this step. To build the model, we used the training data and associated groundtruth labels.

The four functional forms we considered for risk analysis were: exponential fit, power law fit, logistic fit and kernel regression. We discarded exponential or power law fits to model probability in this track. Both are monotonically decreasing functions with a continuously increasing first derivative. The perfect ranking on the training data in this track was better represented a step-like function. Accordingly, the obvious functional form to use here was a logistic function derived using logistic regression, which inherently has a step-like shape.

Like the EM-only-track and the Combined-track, this track also produced a perfect ranking on the training data. So we had numeric issues on this track similar to the ones described for the EM-only-track in Section 6.10.1. We solved those numeric issues in the manner described in that section.

Having solved the numeric issues, we then performed standard logistic regression, which optimizes two parameters in the functional form shown in Equation 4. The following values were derived for these two parameters:

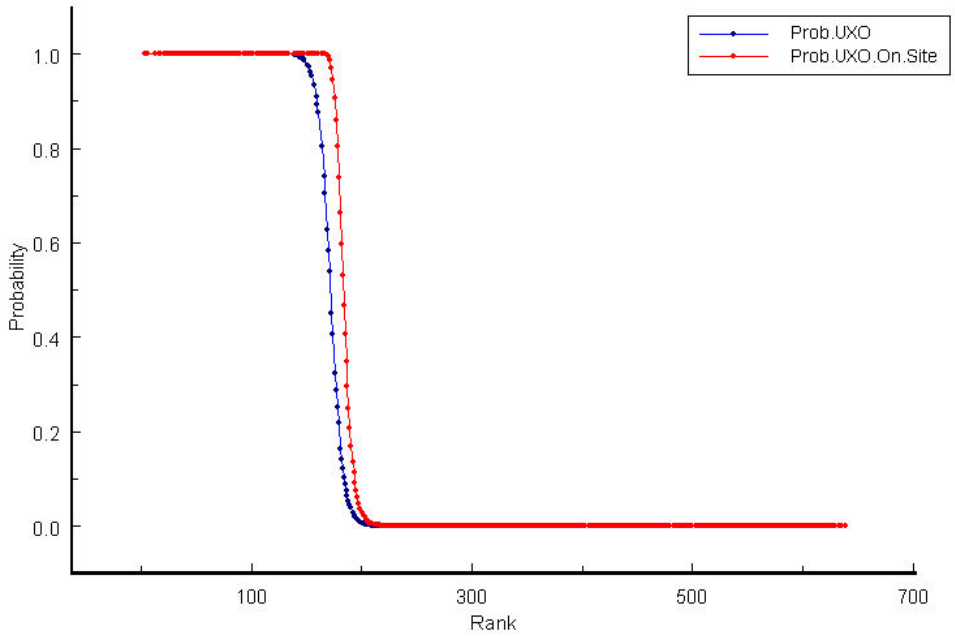
$$\alpha = 30.7149$$

$$\beta = -0.1787$$

Then, we substituted these parameter values into Equation 5 to predict probabilities of UXO on the blind data by rank, using the ranks derived from the blind LGP ensemble predictor scores as the independent variable. These probabilities are shown in the blue line in Figure 65 for the blind targets remaining at this point in the Inversion-track.

Once we derived these probabilities for each blind target, we calculated for each rank, the cumulative probability that one-or-more of the blind targets that have a higher ranking than the rank for which we are making the calculation contain UXO. Those cumulative probabilities are calculated using the “or-of-probabilities” approach described in Equation 2 in Section 2.1.6. These cumulative probabilities that UXO remains on the site are shown in the red line in Figure 65 for the blind targets remaining at this point in the Combined-track.

Figure 65. Residual Risk Analysis for LGP Models on Inversion-track



When the red line falls below a critical *p* value, we assess all targets remaining to the right of that value as high-probability Not-UXO.

The critical value we used was the Bonferonni corrected p-value for a 95% confidence level. We must use the corrected value because we are using three discriminators on this track.[32] Properly corrected, the critical value here is $p \leq 0.01667$. Accordingly, all targets with $p > 0.01667$ were designated as being above the stop-digging threshold; otherwise, below.

8.8 PRIORITIZED DIG-LIST PREPARATION

At this point, we had four sets of targets that needed to be combined into a single prioritized dig-list:

1. Cannot-Analyze Targets
2. Targets excluded as high-probability Not-UXO with the EM_Fit_Coherence pre-discriminator;
3. Targets excluded as high-probability Not-UXO using the EM_Fit_Size pre-discriminator; and
4. The ranked targets from the LGP ensemble predictor.

In the experimental plan, cannot-analyze targets go at the bottom of the prioritized dig-list. Targets that are ranked by the LGP Discriminator as above the stop-digging threshold appear at the top of the list. Three sets of targets should appear below the stop-digging threshold:

[32] See: http://mathworld.wolfram.com/BonferroniCorrection.html.

1. Targets excluded as high-probability Not-UXO with the EM_Fit_Coherence pre-discriminator;
2. Targets excluded as high-probability Not-UXO using the EM_Fit_Size pre-discriminator; and
3. The targets ranked by the LGP Discriminator as below the stop-digging threshold.

These three sets of targets were combined using a $P(UXO)$ generated by residual risk analysis. For items 1 and 2, the $P(UXO)$ used was the $P(UXO)$ generated by the residual risk analysis that excluded the target. For targets described in 3, the $P(UXO)$ used was the value generated by the residual risk analysis on the LGP-generated scores.

8.9 FURTHER ITERATIONS

Because of the high quality of the results produced in the first iteration (reported above), the ESTCP Program Office suggested that we not perform any more iterations and we agreed with that conclusion on the ground that the classification portion of the ROC curve could not be improved in a statistically significant manner, even with more ground-truth.

9 PERFORMANCE ASSESSMENT

9.1 EM-ONLY-TRACK

There were three objectives, each of which is addressed below.

9.1.1 Target of Interest Retention

After we submitted our dig-list on the blind data, the program office scored it and returned the results. Figure 66 shows the ROC chart prepared by the program office for our dig-list on the EM-only track. It should be read as follows: (1) The thick black line on the left side of the chart highlights the 29 cannot-analyze targets, which were dug first; (2) The pink circle identifies the first Not-UXO on our dig-list; (3) The light blue dot represents the last UXO on our prioritized dig-list; (4) The blue dot represents our stop-digging threshold; and (5) The red dots each represent a UXO that was found.

Figure 66. ROC chart showing blind scoring for EM-only-track.

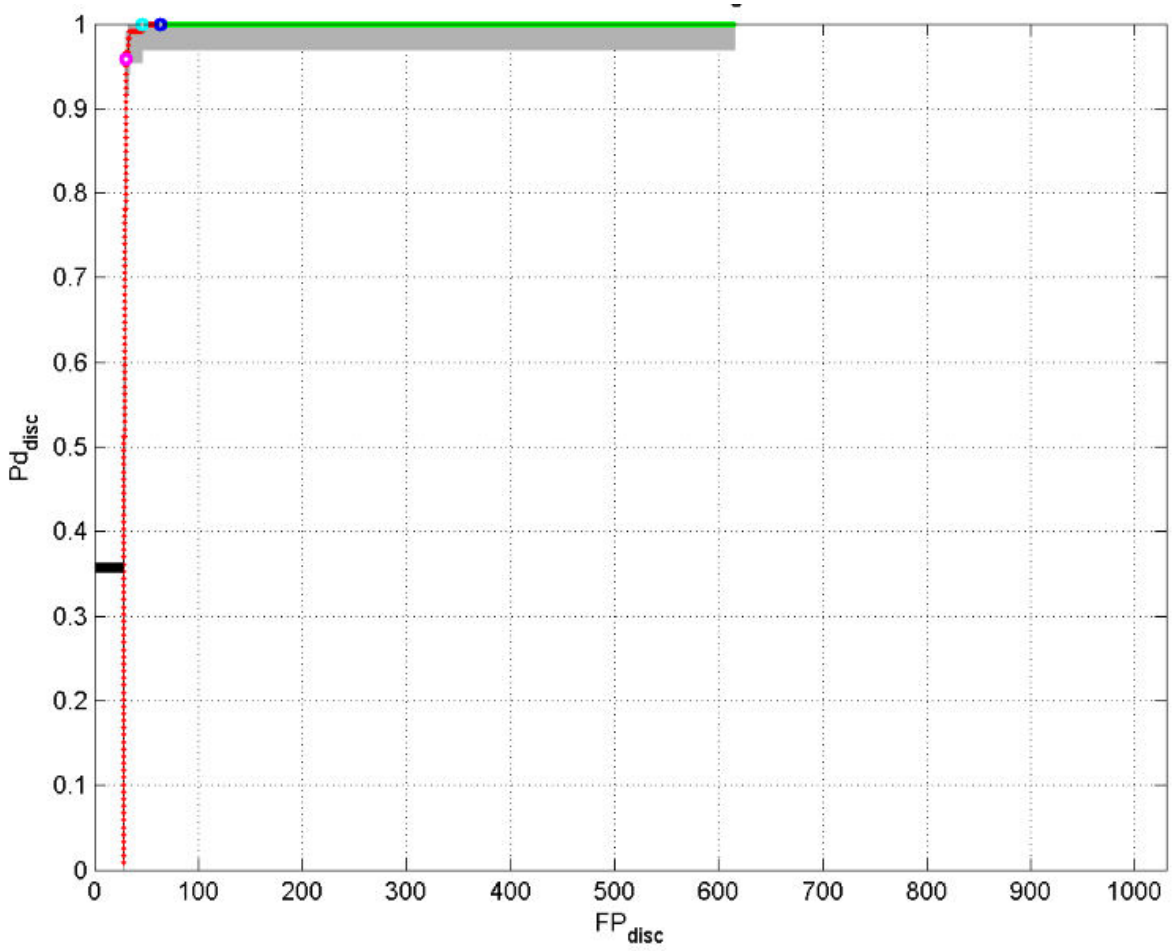


Figure 66 shows that *all* Targets of Interest were retained above our stop-digging threshold. Therefore, this track was a success on this metric.

As noted above, the black line on the left of Figure 66 highlights the cannot-analyze targets. Approximately 4% of the blind targets (twenty-nine targets) were classified as cannot-analyze.

Once we started classifying targets (the near-vertical red line that starts at about FP=29), we generated a near-perfect ROC chart—that is, almost all UXO were ranked above all non-UXO.

The light blue circle shows the final UXO item prioritized on our Inversion-track dig-list. The dark blue circle shows our stop-digging threshold. The key point to draw from these two data is that all UXO were above the stop-digging threshold. That is, no UXO were left in the ground.

Some other observations are appropriate here about track performance. The area under the curve for the ROC curve (counting the cannot-analyze targets) on this track was 0.953.

The area under the curve for the ROC curve (counting only those targets we classified and not-including the cannot-analyze targets) on this track is 0.998. Earlier, given our training data and the LGP models, we estimated that the AUC on the blind data would be 1.0 and the error (1-AUC) would be zero. A blind target AUC of 0.998 and this earlier estimated value of 1.0 are statistically indistinguishable from each other at the 95% confidence level on these data.

There are four conclusions to draw from Figure 66 and the remainder of the EM-only-track section:

1. For the targets it was given to classify, LGP did extremely well, generating an almost perfect classification. Accordingly this track was a success under this objective.
2. Our residual risk analysis correctly determined when it was safe to stop-digging UXO on this track;
3. The combination of LGP Discrimination and Residual Risk Analysis allowed 86.8% of the non-UXO in the study to remain safely in the ground as high probability Not-UXO.
4. With careful modeling, the actual performance on blind UXO data may be closely approximated by the estimated error from even a small training data set. That is, we had already closely estimated the AUC on the blind data when we had completed our models on the training data. That estimate was, within statistical error, a correct estimate.

9.1.2 Non-Target of Interest Reduction

The target for Non-Target of Interest Reduction was that at least 40% of Not-UXO items were left in the ground as high probability Not-UXO. In fact, on this track, we left 89.6% of the Not-UXO in the ground—that is, they were ranked below our stop-digging threshold.

Accordingly, this track was a success on this objective.

9.1.3 Analyze Time and Cost

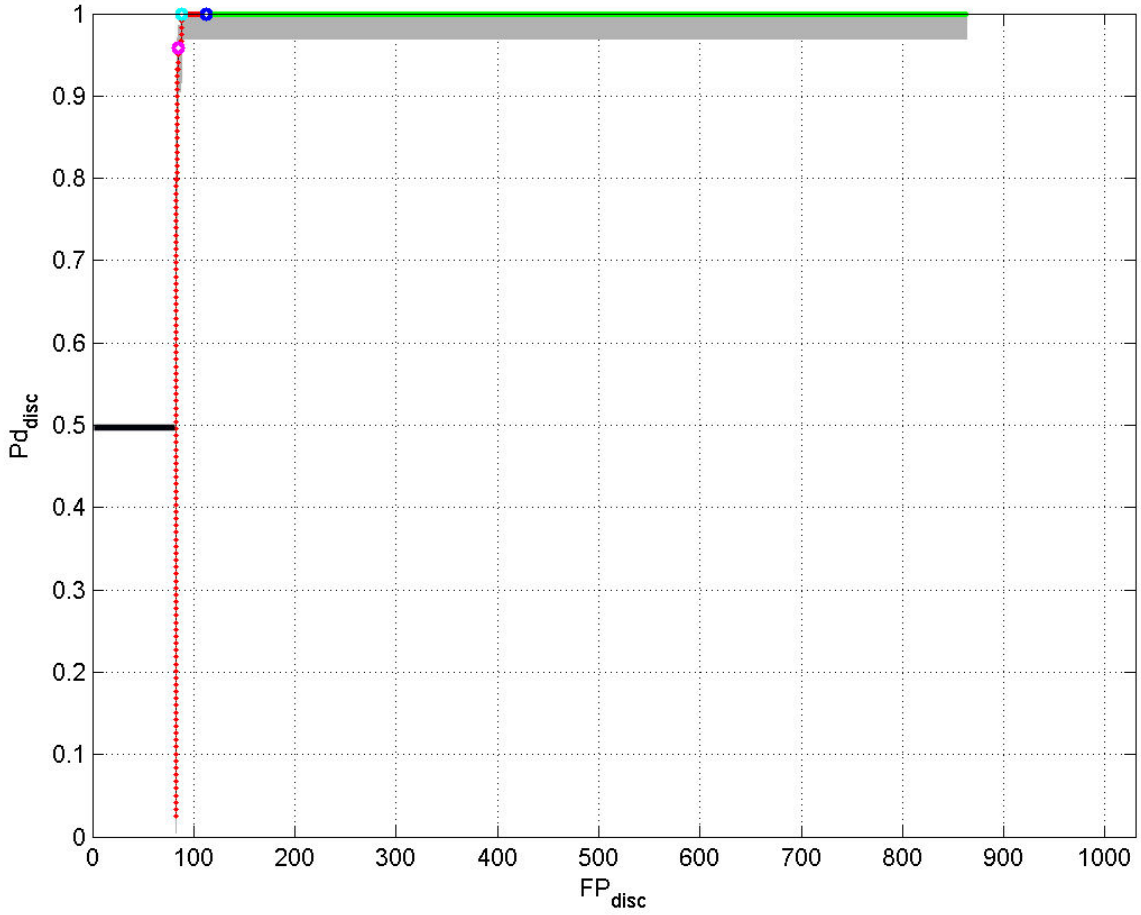
See Section 9.4.

9.2 COMBINED-TRACK

9.2.1 Target of Interest Retention

After we submitted our dig-list on the blind data, the program office scored it and returned the results. Figure 67 shows the ROC chart prepared by the program office from our blind target rankings on the Combined-track. It should be read as follows: (1) The thick black line on the left side of the chart highlights the 86 cannot-analyze targets, which were dug first; (2) The pink circle identifies the first Not-UXO on our dig-list; (3) The light blue dot represents the last UXO on our prioritized dig-list; (4) The blue dot represents our stop-digging threshold; and (5) The red dots each represent a UXO that was found.

Figure 67. ROC chart showing blind scoring for Combined-track.



As noted above, the black line on the left of Figure 67 highlights the cannot-analyze targets. Approximately 7% of all blind targets (86 targets) were classified as cannot-analyze.

Once we started classifying targets (the near-vertical red line that starts at about FP=86), we generated a near-perfect ROC chart—that is, almost all UXO were ranked above all Not-UXO.

The light blue circle shows the final UXO item prioritized on our Inversion-track dig-list. The dark blue circle shows our stop-digging threshold. The key point to draw from these two data is that all UXO were above the stop-digging threshold. That is, no UXO were left in the ground.

Some other observations are appropriate here about track performance. The area under the curve for the ROC curve (counting the cannot-analyze targets) on this track was 0.9035.

The area under the curve for the ROC curve (counting only those targets we classified and not-including the cannot-analyze targets) on this track is 0.999. Earlier, given our training data and the LGP models, we estimated that the AUC on the blind data would be 1.0 and the error (1-AUC) would be zero. A blind target AUC of 0.999 and this earlier estimated value of 1.0 are statistically indistinguishable from each other at the 95% confidence level on these data.

There are six conclusions to draw from Figure 67 and the remainder of the Combined-track section:

1. For the targets it was given to classify, the LGP Discrimination Process did very well, generating an almost perfect classification. Accordingly this track was a success under this objective.
2. Our residual risk analysis correctly determined when it was safe to stop-digging UXO on this track;
3. The combination of LGP Discrimination and Residual Risk Analysis allowed 86.8% of the non-UXO in the study to remain safely in the ground as high probability non-UXO.
4. With careful modeling, the actual performance on blind UXO data may be closely approximated by the estimated error from even a small training data set. That is, we had already closely estimated the AUC on the blind data when we had completed our models on the training data. That estimate was, within statistical error, a correct estimate.
5. The addition of MAG targets to the EM targets on this track resulted in a substantially longer target list and no significant improvement in the quality of the discrimination ROC chart produced. The vast bulk the new MAG targets that were NOT also EM Targets were either very small metal items or nothing at all. Although our pre-discriminator excluded the bulk of these new targets as Not-UXO, the result was an increase in the number of cannot-analyze targets. So while our ROC curve on this track was very good, once we got past the cannot-analyze targets and into LGP classification, this track did not perform as well as the EM-only-track because of the increased number of cannot-analyze targets.

9.2.2 Non-Target of Interest Reduction

The target for Non-Target of Interest Reduction was that at least 40% of Not-UXO items were left in the ground as high probability Not-UXO. In fact, on this track, we left 86.8% of the Not-UXO in the ground—that is, they were ranked below our stop-digging threshold.

Accordingly, this track was a success on this objective.

9.2.3 Analyze Time and Cost

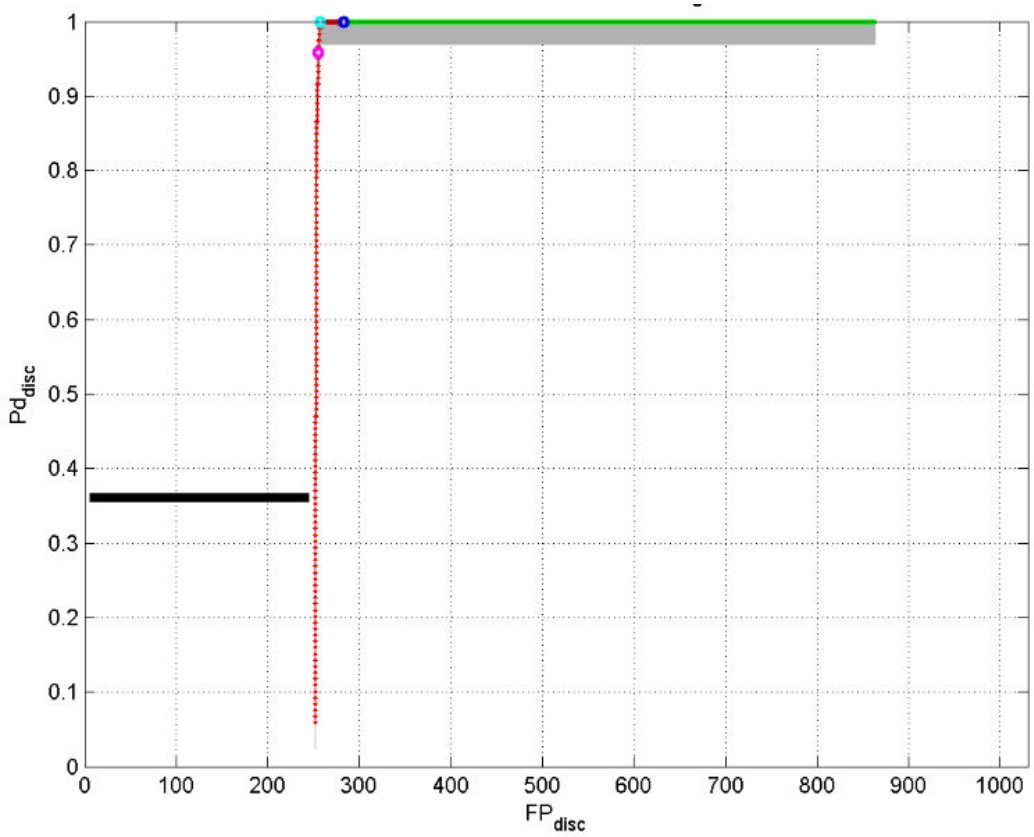
See Section 9.4.

9.3 INVERSION-TRACK

9.3.1 Target of Interest Retention

After we submitted our dig-list on the blind data, the program office scored it and returned the results. Figure 68 shows the ROC chart prepared by the program office for our scoring.

Figure 68. ROC chart showing blind scoring for Inversion-track.



The black line on the left highlights the cannot-analyze targets for this track. Approximately 26% of all blind targets (260 targets) were classified as cannot-analyze.

Once we started classifying targets (the near-vertical red line that starts at about FP=260), we generated a near-perfect ROC chart—that is, almost all UXO were ranked above all non-UXO.

The light blue circle shows the final UXO item prioritized on our Inversion-track dig-list. The dark blue circle shows our stop-digging threshold. The key point to draw from these two data is that all UXO were above the stop-digging threshold. That is, no UXO were left in the ground.

Therefore, this track was a success on this objective, which was 100% retention of Targets of Interest (UXO).

Some other observations are appropriate here about track performance.

The area under the curve for the ROC curve (counting the cannot-analyze targets) on this track was 0.715.

The area under the curve for the ROC curve (counting only those targets we classified and not-including the cannot-analyze targets) on this track is 0.999. Earlier, given our training data and the LGP models, we estimated that the AUC on the blind data would be 1.0 and the error (1-AUC) would be zero. A blind target AUC of 0.999 and this earlier estimated value of 1are statistically indistinguishable from each other at the 95% confidence level on these data.

There are six conclusions to draw from Figure 68 and the and the data that supported it:

1. The track objective was met;
2. The primary purpose of this track was to assess LGP as a classifier. For the targets it was given to classify, LGP did extremely well, generating an almost perfect classification;
3. Our residual risk analysis correctly determined when it was safe to stop-digging UXO on this track, notwithstanding the relatively high number of cannot-analyze targets;
4. The combination of LGP Discrimination and Residual Risk Analysis allowed 67% of the non-UXO in the study to remain safely in the ground as high probability non-UXO.
5. With careful modeling, the actual performance on blind UXO data may be closely approximated by the estimated error from even a small training data set. That is, we had already closely estimated the AUC on the blind data when we had completed our models on the training data. That estimate was, within statistical error, a correct estimate.
6. For classification using inversion-based attributes, in order to have a reasonable number of cannot-analyze targets, it is necessary to tolerate inversions that produce very imperfect coherence results. This section demonstrates a principled and statistically valid way to reduce the number of cannot-analyze targets and still maintain high modeling standards. That said, we were unable to reduce the number of cannot-analyze targets to a range competitive with the EM-only and the Combined-tracks, even using these techniques.

9.3.2 Non-Target of Interest Reduction

The target objective for Non-Target of Interest Reduction was that at least 40% of Not-UXO items were left in the ground as high probability Not-UXO. In fact, on this track, we left 67.1% of the Not-UXO in the ground—that is, they were ranked below our stop-digging threshold.

Accordingly, this track was a success on this objective.

9.3.3 Analyze Time and Cost

See Section 9.4.

9.4 Time and Cost Analysis

The target for Time and Cost is that no more than 60 man-days of time would be spent in analysis before the stop-digging threshold was set. We set three stop digging thresholds, one for each track. Accordingly, we break this objective down by track.

9.4.1 EM-Only-Track

We spent 74.5 man-days in production on the EM-only-track. This exceeded the objective. This occurred for two reasons: (1) We were establishing procedures and processes on this track and there was a good deal of backtracking to make sure we had a good trace on the process that produced the results; (2) The rut-noise discussed elsewhere in this report forced us to change our process for ellipse extraction on this track. Although this figure does not include time spent trying to solve the rut-noise problem in what turned out to be unproductive ways, a good deal of the time spent addressing the rut-noise is fairly allocable to our production time. This track is the first time we addressed the rut noise; accordingly, it occupied a good deal more time than it did on the Combined-track, where it was a familiar problem.

9.4.2 Combined-track

We spent 52 man-days in production on the Combined-track. That met our objective.

9.4.3 Inversion-Track

We spent 23 man-days in production on the Inversion-track. That met our objective.

10 CONCLUSION

This study strongly supports the conclusion that the LGP Discrimination Process™ performs highly statistically significant discrimination on large ordnance items in a manner that would greatly reduce the number of digs necessary to clear a site containing such items.